UNIVERSITÀ DEGLI STUDI DI MILANO

Facoltà di Scienze Matematiche, Fisiche e Naturali

Dipartimento di Tecnologie dell'Informazione

Corso di Laurea in Scienze e Tecnologie dell'Informazione

METODI DI OTTIMIZZAZIONE DEL LIVELLO DI SERVIZIO DEL SISTEMA "118" DI MILANO

RELATORE:

Prof. Giovanni RIGHINI

CORRELATORE:

Dott. Giovanni SESANA

TESI DI LAUREA DI:

Federico FICARELLI

Matricola: 703467

Anno Accademico 2006 / 2007

Indice

Capitolo 1

Introduzione

1.1 Panoramica del Servizio Sanitario Unità di Emergenza (S.S.U.Em.) 118 di Milano

Il Servizio “118” viene fatto comunemente corrispondere al numero telefonico universale che mette in comunicazione il cittadino con le centrali operative del sistema di allarme sanitario. Per quanto riguarda la centrale oggetto di analisi in questo lavoro di tesi, essa è dislocata presso l’Azienda Ospedaliera Niguarda Ca’ Granda di Milano e serve qualsiasi richiesta provenga dal territorio di tutta la provincia. Il bacino d’utenza si aggira intorno ai 3 milioni e 720 mila abitanti e, fra le dodici presenti in Lombardia, quella di Milano è la centrale di maggiori dimensioni con un’area di competenza di circa 1440 km^2 e 118 comuni. In media si hanno 1580 contatti giornalieri e di questi, il 42% richiede un intervento di soccorso; annualmente si ha una media di 650 mila chiamate e 252 mila soccorsi effettivi ([25]).

Il Servizio “118” si presenta come un sistema estremamente complesso, in cui molti decisori hanno voce in capitolo riguardo l’organizzazione, lavora nell’incertezza ed è tenuto a garantire un livello di qualità della risposta in termini di tempi di intervento, definito a livello normativo. È da considerare che, a differenza di quello che comunemente si pensa, il Servizio “118” non corrisponde unicamente al recapito telefonico ma è un sistema distribuito su tutto il territorio di competenza, composto di risorse eterogenee, umane e materiali, infrastrutture e procedure da seguire rigidamente.

La risorsa primaria a disposizione del servizio sono le ambulanze, in parte di proprietà dell’Azienda Ospedaliera, in parte facenti capo ad associazioni di volontariato sostanzialmente private. Le ambulanze *di proprietà* sono 52, di cui 24 normalmente a disposizione per l’area urbana di Milano e le restanti 28 dislocate nel resto della provincia. Inoltre, sono a disposizione circa 100 mezzi di proprietà delle associazioni di volontariato sotto contratto di convenzione. Questi mezzi, detti *“a gettone”*, sono prenotabili dalla centrale operativa che ne allerta gli equipaggi per determinati periodi di tempo. Questi mezzi possono essere liberamente dislocati sul territorio e la loro tariffa viene pagata solamente all’atto di un effettivo impiego in missione. Le altre tipologie di mezzi a disposizione sono una decina di auto mediche ed un elicottero. L’intero sistema è coordinato dalla centrale di Milano presso la quale sono impiegati 15 medici, 14 infermieri e 60 operatori tecnici. Degli equipaggi fanno parte a rotazione 200 medici e 300 infermieri mentre ammonta a circa un migliaio il numero di volontari coinvolti nelle associazioni private.

La procedura di risposta seguita dagli operatori della centrale è composta dalle seguenti attività:

1. all'arrivo una richiesta di soccorso si procede alla valutazione della tipologia e gravità. È compito dell'operatore effettuare la valutazione tentando di raccogliere le informazioni necessarie per assegnare un codice, universalmente noto come *Triage*, all'intervento. I codici, associati ad un colore ed ordinati per gravità crescente ([26]), sono:

   bianco : nessuna criticità, assenza di rischi e traumi;

   verde : lieve criticità, parametri vitali nella norma, assenza di rischi evolutivi, trauma minore;

   giallo : media criticità, alterazione di un solo parametro vitale, trauma senza fattori aggravanti;

   rosso : elevata criticità, alterazione di almeno due parametri vitali, assenza di un parametro vitale.

   Ai fini della trattazione, è importante segnalare che in caso di codice verde le ambulanze circolano senza sirena accesa e devono di conseguenza rispettare il codice della strada come un qualsiasi mezzo. Solo nel caso di missione con codice giallo o rosso è consentito l'utilizzo della sirena e la conseguente circolazione di emergenza;

2. valutata la gravità del caso, l'operatore procede con la raccolta delle informazioni anagrafiche e l'indirizzo del cittadino ed invia un mezzo disponibile che sia idoneo a gestire la natura dell'intervento; il contatto con l'equipaggio viene mantenuto via radio o telefono cellulare;

3. in caso di bisogno di ulteriori cure, viene organizzato il trasporto del paziente in un ospedale attrezzato e pronto per prestare i trattamenti aggiuntivi. L'operatore deve quindi avvertire la struttura di destinazione dell'imminente arrivo del paziente;

4. in situazione di particolare complessità (incidenti di grandi dimensioni, calamità, incendi, ecc…), vengono contattate altre centrali operative (vigili del fuoco, forze dell'ordine, protezione civile);

5. se necessario, l'operatore è tenuto a fornire istruzioni semplici ed essenziali per il primo soccorso da effettuare in attesa dell'arrivo dell'ambulanza.

1.2 Le problematiche affrontate

Nell'ambito della gestione del Servizio "118", il decisore si trova quotidianamente di fronte ad una serie di problematiche relative al dimensionamento delle risorse, alla loro organizzazione sia tattica che strategica, problematiche relative alla fruizione da parte degli utenti ed al miglioramento dell'efficienza complessiva. Attualmente il processo decisionale è governato dall'esperienza dell'operatore che, avendo spesso anni di servizio all'attivo, riesce a dominare in modo accettabile la complessità delle dinamiche del sistema. Il fatto che le prestazioni ed il corretto funzionamento dipendano unicamente dalle capacità del singolo e dalle sue decisioni dettate dall'esperienza comincia a mostrare chiari segni di inadeguatezza. Come segnalato dai responsabili della centrale operativa di Milano, gli aspetti che risultano critici sono i seguenti:

- le competenze, maturate grazie all'esperienza sul campo, rimangono patrimonio del singolo e non sono facilmente trasferibili. Spesso l'operatore stesso non è in grado di spiegare quali sono i criteri decisionali che intervengono nelle sue scelte. Questo aspetto si presenta sia a livello tattico che strategico.

- Manca qualsiasi strumento che consenta di elaborare decisioni di gruppo. Il decisore ha espresso la necessità di avere a disposizione strumenti di supporto che facilitino i processi decisionali di gruppo, permettendo di valutare in modo formale le proposte elaborate, analizzarne i punti deboli e facendo sì che questa conoscenza sia costruita collettivamente.

- Spesso l'esperienza non è sufficiente: è assodato che in fase di dimensionamento delle risorse, di fronte ad un minimo sospetto di situazione critica il decisore tende ad assumere un comportamento atto ad "andare sul sicuro", spesso sovradimensionandone enormemente il fabbisogno. Questo comportamento è molto usato proprio perché risulta essere l'unico strumento a disposizione per evitare un pericoloso sottodimensionamento.

- Assume una discreta importanza anche l'impatto che uno strumento formale avrebbe sull'aspetto politico del dimensionamento. Se il decisore fosse in possesso di analisi prodotte da uno strumento di supporto, le richieste di aumento delle risorse presso le autorità finanziatrici (in questo caso, la Regione Lombardia) sarebbero supportate da dati concreti e quindi maggiormente comprese.

Le problematiche che il decisore si trova ad affrontare e per le quali ha richiesto un supporto riguardano, come già accennato, sia aspetti strategici che tattici. Un primo processo è quello che riguarda il dimensionamento delle risorse: con una data cadenza, tipicamente ogni giorno, il decisore si trova a dover stimare di quali risorse avrà bisogno per far fronte alla situazione futura, generalmente per il giorno successivo. Attualmente questo processo avviene basandosi sull'esperienza: egli è in grado di valutare il periodo dell'anno, l'incidenza delle condizioni atmosferiche, la stagionalità, le festività, gli eventi straordinari e, combinando tutte queste valutazioni, stimare di quanti mezzi si avrà bisogno per garantire un servizio accettabile. Se c'è necessità di un'integrazione della flotta di proprietà, verrà inoltrata una richiesta per mobilitare quanti mezzi a noleggio si ritengono sufficienti. Altri aspetti sui quali il decisore è chiamato a intervenire è la "tolleranza" d'attesa delle richieste non urgenti. In situazioni particolarmente critiche infatti, viene stabilita una soglia temporale entro la quale i pazienti che non sono in pericolo di vita possono essere lasciati in attesa per dare priorità immediata ai casi più gravi. La durata di questo periodo "cuscinetto" è un'altra leva sulla quale il decisore può agire per alleggerire il carico o migliorare la risposta del servizio alle urgenze.

Un aspetto squisitamente tattico riguarda tutte le problematiche geografiche legate alle risorse. Il decisore è chiamato in causa anche nel momento in cui sia necessario decidere dove realizzare i punti d'attesa, le "colonnine" dove le ambulanze attendono l'assegnamento di una missione. La capacità di costruire questi punti d'attesa è limitata ed il loro posizionamento sul territorio può influire pesantemente sulle prestazioni complessive del servizio. Un altro aspetto legato a queste problematiche è l'ottimizzazione della copertura del territorio, ovvero dove posizionare i mezzi disponibili per fare in modo di coprire il maggior numero di aree anche in base alla loro importanza (densità di popolazione, criticità, ecc...). La disposizione dei mezzi può inoltre variare a seconda del numero di ambulanze libere in un dato istante: nasce quindi il problema di ottimizzare la copertura del servizio, limitando nel contempo gli spostamenti dei mezzi effettuati senza l'assegnazione di una missione, solamente allo scopo di ricollocarsi presso punti d'attesa migliori.

1.3 Strutturazione del lavoro di tesi

La prima parte del lavoro di tesi ha previsto l'analisi dei dati e la definizione delle procedure per la loro elaborazione. Il Capitolo 2 presenta la struttura dei dati storici forniti dalla centrale operativa, il registro relativo a tutti gli eventi avvenuti in un intero anno di funzionamento del Servizio "118"; nella stessa sezione vengono descritti i dati geografici ed il loro utilizzo tramite un Sistema Informativo Geografico. La definizione delle procedure di elaborazione dati ha richiesto la realizzazione di alcuni strumenti come un algoritmo per il cammino minimo su rete stradale, un algoritmo per il geocoding robusto ed un modello di programmazione matematica per la suddivisione in fasce temporali del periodo di attività.

Nell'ambito del Capitolo 3 viene affrontata la realizzazione di modelli in grado di fornire un utile supporto alla fase decisionale di dimensionamento del servizio. L'idea guida è stata quella di arrivare a produrre le informazioni di cui il decisore ha bisogno per migliorare quei processi che fino ad ora sono stati governati unicamente dall'esperienza maturata dal singolo. A tale scopo sono stati realizzati dei modelli di sistemi a coda che permettono, a fronte di una data configurazione delle variabili decisionali d'interesse, di valutare il livello di servizio del sistema, le sue prestazioni e punti deboli e decidere eventuali azioni correttive sul fabbisogno di risorse.

Il Capitolo 4 riporta il lavoro svolto per realizzare modelli di ottimizzazione della copertura del territorio. Dopo una revisione della letteratura riguardante l'argomento e la rassegna dei risultati ottenuti in questo campo, la trattazione presenta la definizione di tre varianti del classico problema di *set covering* adattato per considerare specifiche necessità di servizio. In questa sezione si sono tenute in particolare considerazione le esigenze espresse dal decisore: l'obiettivo è stato infatti quello di fornire una serie di modelli che permettessero l'ottimizzazione del servizio sul territorio, aiutando la soluzione delle problematiche geografiche come il posizionamento dei mezzi, la gestione delle loro rilocalizzazioni, la costruzione dei punti di stazionamento sull'area di competenza.

Capitolo 2

Elaborazione dei dati storici e geografici

Introduzione

L'obiettivo del lavoro di tesi è quello di realizzare una serie di modelli e metodologie che fungano da strumento per il supporto alle decisioni nell'ambito della gestione del Servizio "118". Per raggiungere con profitto questo scopo, risulta di particolare importanza la fase preliminare dedicata all'acquisizione ed analisi dei dati. Le informazioni, tutte fornite dalla centrale operativa dell'ospedale Niguarda Ca' Granda di Milano, sono le stesse con le quali gli operatori lavorano quotidianamente nella loro attività di gestione delle emergenze. La loro analisi ha permesso di integrare i modelli sviluppati nel seguito del lavoro con dati reali, acquisiti direttamente sul campo, valutandone il comportamento così come se si trovassero ad operare già presso la centrale.

Le informazioni sono di due tipologie differenti: la prima è lo storico di servizio che, analizzato nel Paragrafo 2.1, riporta il registro dell'attività svolta dal Servizio "118" durante tutto l'anno 2005. I dati riguardano gli interventi di soccorso, le tratte percorse dai mezzi, le caratteristiche dei punti di stazionamento e tutti quegli aspetti necessari alla ricostruzione del funzionamento del sistema. La seconda categoria di dati riguarda l'aspetto geografico, direttamente legato all'area di competenza. Grazie alle mappe della rete stradale ed all'utilizzo di un Sistema Informativo Geografico, è stata realizzata un'analisi di queste informazioni e, come spiegato nel Paragrafo 2.2, è stata costruita una struttura dati che modella la rete stradale di tutta l'area di competenza. La disponibilità di dati geografici ha consentito la localizzazione di tutti gli eventi riportati nello storico di attività, legando così l'aspetto temporale a quello spaziale del territorio. Nell'ultima parte del capitolo, il Paragrafo 2.3, sono presentati gli strumenti realizzati per arrivare a definire una serie di procedure di elaborazione; tramite queste metodologie è possibile calcolare tutti i dati necessari alla definizione dei modelli che verranno presentati nel seguito del lavoro di tesi.

2.1 Formato dei dati storici di servizio

La prima fase del lavoro ha previsto l'analisi dei dati storici di attività del servizio. La centrale operativa ha fornito un registro relativo a tutte le missioni svolte nell'anno 2005, organizzato in un insieme di file testuali, ognuno riguardante un aspetto del funzionamento del servizio.

In Tabella 2.1 è riportata la struttura del registro delle missioni svolte. In questo schema sono presenti tutti i campi necessari alla caratterizzazione di una singola missione, dall'iden-

STORICO MISSIONI		
Campo	**Formato**	**Descrizione**
ID_MISSIONE	50000012	Identificativo univoco missione
DT_MISSIONE	01/01/05 00.26.00	Data ed ora dei inizio missione
ID_SOCCORSO	50000001	Identificativo intervento di soccorso
CD_MEZZO	VNO007	Identificativo univoco mezzo (targa)
DS_TP_MEZZO	AMBULANZA TIPO A	Tipologia di mezzo impiegato
DS_CONVENZ	GET	Tipologia di convenzione del mezzo
ID_CODICE	G	Codice urgenza assegnato in automatico dal sistema
ID_CODICE_E	G	Codice urgenza assegnato dall'operatore
VL_TMPMIS	2286	Tempo in secondi occorsi per lo svolgimento dell'intera missione
DT_FINE_MI	01/01/05 01.06.00	Data ed ora di fine missione
DT_INIZIO_R	01/01/05 01.05.00	Data ed ora di uscita dal pronto soccorso/colonnina
DS_LG_DEST	CORNAREDO VERDE S.PIETRO	Luogo di destinazione
DS_PUNSTA	MACIACHINI	Punto di stazionamento (colonnina)
ID_ENTE_PT	1156	Codice identificativo ente
ID_MISS_INT	50000001	Codice missione (eventualmente) interrotta in favore della corrente
ID_CODICE_T	V	Codice urgenza assegnato al trasporto

Tabella 2.1: formato e descrizione dei campi per lo storico missioni.

tificativo univoco agli istanti di inizio e di fine. Alcune informazioni sono risultate di grande importanza per l'analisi del funzionamento del servizio: il campo DS_CONVENZ ad esempio, indica il tipo di convenzione alla quale è sottoposta il mezzo impegnato in missione, se di proprietà o "a gettone" mentre DS_PUNSTA indica presso quale punto d'attesa il mezzo era prima dell'assegnazione della missione. Per quanto riguarda il codice di urgenza, i campi sono molteplici. Il primo, ID_CODICE, indica il livello di emergenza con il quale è stata classificata la missione da parte del sistema automatico; il secondo, ID_CODICE_E, è il codice di urgenza assegnato dall'operatore che ha ricevuto la chiamata ed è con questo livello di priorità che l'ambulanza può circolare. Per l'anno 2005, i dati a disposizione riportano un numero complessivo di missioni svolte pari a 247596.

PUNTI DI STAZIONAMENTO		
Campo	**Formato**	**Descrizione**
ID_PUNSTA	12	Identificativo univoco colonnina
DS_PUNSTA	MACIACHINI	Nome punto di stazionamento
DS_VIA_1	MACIACHINI CARLO	Via del punto di stazionamento
VL_RIF_X	1514601	Coordinata geografica x (Gauss-Boaga)
VL_RIF_Y	5038243	Coordinata geografica y (Gauss-Boaga)
ID_LOCALITA	1	Identificativo località
DS_TP_PUNSTA	COLONNINA	Tipologia del punto di stazionamento

Tabella 2.2: formato e descrizione dei campi per i dati riguardanti i punti di stazionamento.

In Tabella 2.2 è riportato lo schema dei dati relativi ai punti di stazionamento presenti sul territorio. I campi presenti permettono di caratterizzare ogni punto d'attesa, dalla sua tipologia (campo DS_TP_PUNSTA) alle sue coordinate geografiche (campi VL_RIF_X e VL_RIF_Y). Queste ultime in particolare, saranno di grande importanza in fase di analisi dei dati geografici (Paragrafo 2.2) in quanto ne permettono la localizzazione sul territorio.

In Tabella 2.3 è riportato lo schema dei dati relativi alle tratte percorse dai mezzi di soccorso. I dati presenti in questo registro sono di particolare importanza: i campi consentono di collegare i registri riguardanti missioni, punti di stazionamento ed interventi di soccorso permettendo di ricostruire sia informazioni temporali (durata) che spaziali (distanza percorsa) di ogni spostamento. In particolare, il campo ID_MISSIONE funge da chiave per collegare la tupla con la corrispondente nella Tabella 2.1, dalla quale si può ricavare il punto di sta-

TRATTE		
Campo	**Formato**	**Descrizione**
ID_TRATTA	42710	Identificativo univoco tratta
DT_ARRIVO	01/01/05 00.09.00	Data ed ora di arrivo sul luogo dell'evento
DT_PARTENZA	01/01/05 00.40.00	Data ed ora di partenza dal luogo dell'evento
DS_LOCALITA	TANGENZIALE OVEST	Località dell'evento
DS_VIA	6 KM SS 11 NOVARA	Via luogo dell'evento
VL_RIF_X	1505200	Coordinata geografica x (Gauss-Boaga)
VL_RIF_Y	5038800	Coordinata geografica y (Gauss-Boaga)
DS_DESTINAZ	INTERVENTO	Tipologia di tratta
ID_MISSIONE	50000012	Identificativo univoco missione
ID_LG_DEST	25000938	Identificativo univoco luogo destinazione
ID_VIA	42710	Identificativo univoco via di destinazione

Tabella 2.3: formato e descrizione dei campi per i dati riguardanti le tratte.

zionamento di partenza e, di conseguenza, le sue coordinate geografiche. Per l'anno 2005, il registro fornito riporta i dati relativi a 439657 tratte percorse.

In Tabella 2.4 è presentato lo schema dei dati riguardanti il registro degli interventi di soccorso effettuati. Le tuple fornite in questa tabella, per un totale di 576568 riferite all'anno 2005, riguardano le caratteristiche degli interventi di soccorso: tipologia, indirizzo, coordinate geografiche. Tramite il campo ID_SOCCORSO di Tabella 2.1, è possibile associare ad ognuno degli interventi di questo registro una missione, dalla quale poi è possibile ricavare tutti i risultati su distanze e tempistiche di percorrenza.

La localizzazione di tutte le chiamate di emergenza che hanno richiesto l'intervento di un'ambulanza, dando origine ad una missione registrata nello storico, è rappresentata in Figura 2.1.

Data la natura dai dati di cui sono composti i registri, ovvero semplici file di testo, si è posto il problema di come organizzare la mole di informazioni a disposizione per renderne più facile l'accesso e l'elaborazione. Si è deciso di importare tutte le tabelle viste in questo paragrafo in un database relazionale che ha assunto lo stesso schema dei dati iniziali. Per completare questa importazione sono stati sfruttati come chiavi quei campi che forniscono una connessione logica fra le tuple dei vari registri.

INTERVENTI DI SOCCORSO		
Campo	**Formato**	**Descrizione**
ID_SOCCORSO	50000001	Identificativo univoco soccorso
DT_SOCCORSO	01/01/05 00.01.00	Data ed ora di inizio soccorso
ID_LOCALITA	205	Identificativo univoco località intervento
DS_LOCALITA	LAINATE	Località dell'intervento
ID_LOC_COM	307	Identificativo del comune di intervento
DS_COMUNE	LAINATE	Nome del comune di intervento
CD_PROVINCIA	MI	Provincia di intervento
ID_VIA_1	2050080	Identificativo univoco via dell'intervento
DS_VIA_1	RIMEMBRANZE	Nome via dell'intervento
ID_VIA_2	2050080	Identificativo univoco eventuale incrocio
DS_VIA_2	RIMEMBRANZE	Nome via eventuale incrocio
VL_RIF_X	1505200	Coordinata geografica x (Gauss-Boaga)
VL_RIF_Y	5038800	Coordinata geografica y (Gauss-Boaga)
ID_CLASSIFICZ	7	Codice della tipologia di intervento
DS_MOTIVO	INC. STRADALE	Motivo dell'intervento
ID_CODICE	G	Codice urgenza
ID_LUOGO	4273410	Identificativo univoco luogo dell'intervento
DT_CHIUSURA	01/01/05 00.01.00	Data ed ora di fine dell'intervento

Tabella 2.4: formato e descrizione dei campi per i dati riguardanti gli interventi di soccorso.

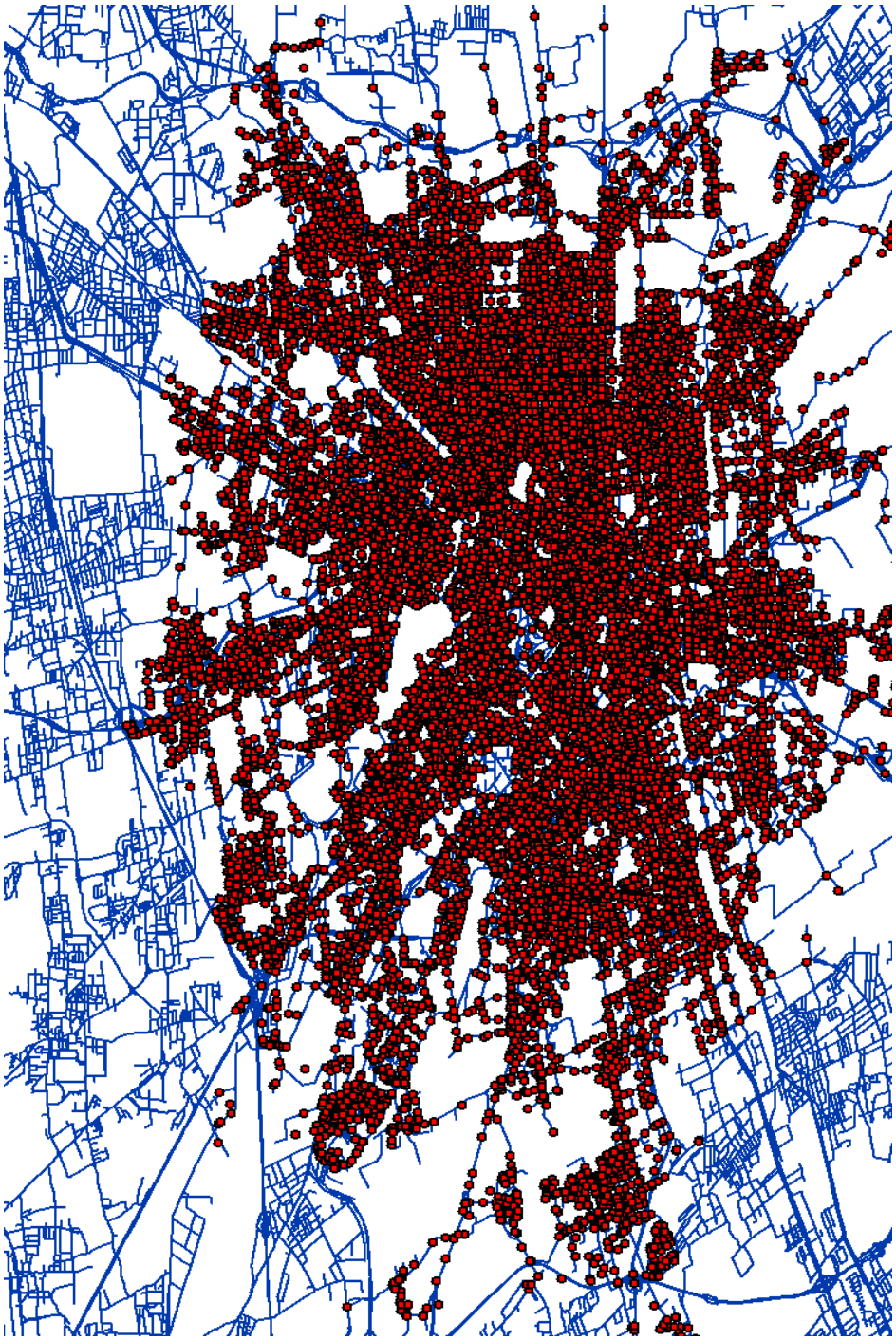


Figura 2.1: localizzazione di tutte le chiamate provenienti dall'area urbana di Milano che hanno richiesto l'intervento di un'ambulanza.

2.2 Analisi dei dati geografici

L'analisi dei dati di servizio ha previsto l'utilizzo di dati geografici, forniti dalla centrale operativa di Milano, riguardanti il territorio di competenza del Servizio "118", rappresentato in Figura 2.2. Le mappe relative al sistema di viabilità permettono di localizzare i dati storici per posizionarli sul territorio, aggiungendo così una dimensione, quella spaziale, all'informazione complessiva. Grazie all'utilizzo di un Sistema Informativo Geografico è stato

Figura 2.2: mappa dell'area di competenza della centrale operativa di Milano.

possibile localizzare ogni evento riportato nello storico presso un luogo reale, calcolare le distanze percorse e valutare l'incidenza delle chiamate di soccorso per ogni zona.

2.2.1 Introduzione ai Sistemi Informativi Geografici

I dati geografici relativi all'area di competenza riguardano la rete stradale e le sue caratteristiche di viabilità. Per lavorare con dati di questa natura si è optato per l'utilizzo di un *Sistema Informativo Geografico* o *GIS*, dall'inglese *Geographical Information System*. I GIS sono una collezione di strumenti software che permettono la gestione, l'elaborazione e la visualizzazione di dati spaziali. Questi sistemi permettono l'associazione fra dati geografici ed

altre informazioni e viceversa, consentendo quindi di *geo-referenziare*, ovvero posizionare sul territorio, qualsiasi tipo di informazione ([4]).

Un tipico sistema geografico permette di lavorare con i dati geografici da tre punti di vista:

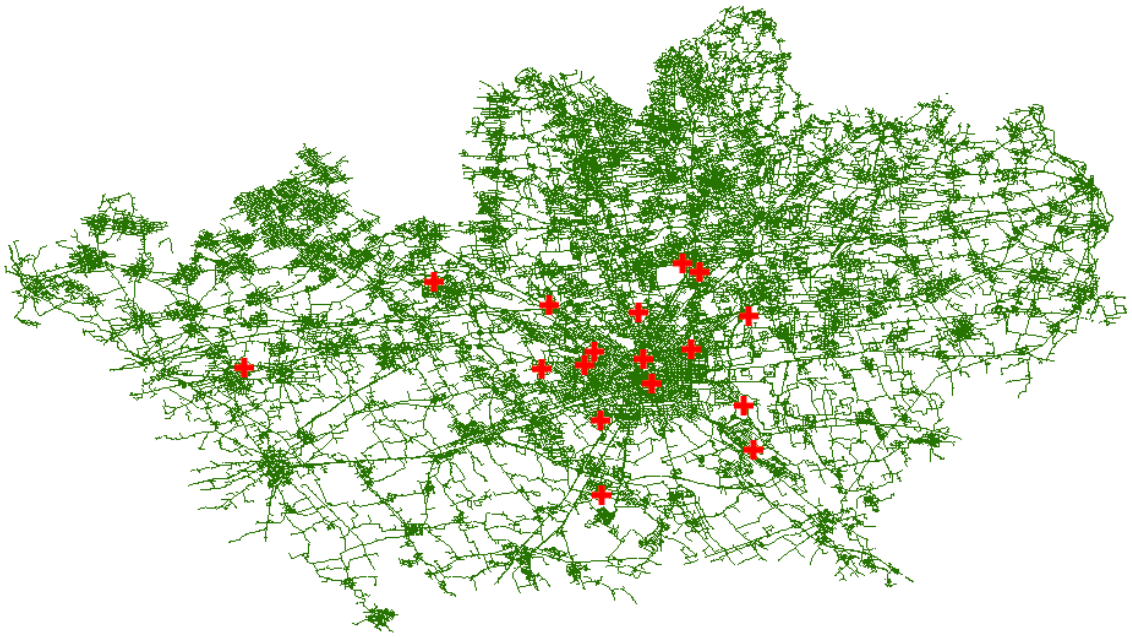
- Base di dati. Da un punto di vista puramente implementativo, le fondamenta di un GIS sono di fatto un database relazionale in grado di gestire informazioni di natura spaziale. Questa struttura basilare viene estesa per permettere la trattazione di dati vettoriali, raster, topografici o cartografici. Un'altra caratteristica comune è la capacità di strutturare i dati geografici in livelli tematici detti *layer*, ovvero raggruppare logicamente insiemi di oggetti omogenei. Nel caso di questo lavoro, la suddivisione in livelli è risultata estremamente utile: grazie a questa caratteristica è stato possibile mantenere come "strato" fondamentale quello relativo alla rete stradale ed aggiungere poi, a seconda delle necessità, layer aggiuntivi come quello relativo agli ospedali, ai punti di stazionamento o di costruire livelli come risultato di elaborazioni.

- Elaborazione. Un GIS è anche un insieme di strumenti che consentono l'analisi e l'elaborazione dei dati geografici. Il supporto al *processing* fornisce anche la possibilità per l'utente di applicare le proprie funzioni, i propri algoritmi ai dati gestiti dal sistema ed eventualmente di memorizzarne i risultati sotto forma di informazioni spaziali. Si veda ad esempio la Figura 2.4: il layer di classificazione dei punti è il risultato di una elaborazione effettuata sul livello base, quello stradale.

- Visualizzazione. Una delle funzioni tipiche di un GIS è quella di permettere la costruzione di rappresentazioni geografiche, anche complesse, nelle quali vengono visualizzati i dati e le loro relazioni spaziali. Le *mappe* sono il principale strumento per presentare l'informazione geografica agli utenti e consentirne l'interazione.

Nell'ambito del lavoro svolto in questa tesi, l'utilizzo di un GIS ha permesso di trattare i dati geografici e stradali relativi alla zona di competenza del servizio, forniti dalla centrale operativa di Milano. L'esportazione da parte del sistema di una serie di interfacce di sviluppo dedicate, ha permesso inoltre l'integrazione degli algoritmi di analisi ed ottimizzazione con le informazioni geografiche, rendendoli in grado di operare direttamente sul territorio.

I dati forniti riguardano la struttura di viabilità sia della provincia che della città di Milano; le mappe utilizzate in questo lavoro sono raffigurate in Figura 2.3.



(a) Area provinciale con strutture ospedaliere.

(b) Area urbana con strutture ospedaliere.

Figura 2.3: mappe GIS della rete stradale di viabilità dell'area di competenza della centrale operativa di Milano.

Il formato utilizzato è il *NavTech* (Navigation Technologies, NAVSTREETS Street Data, Version 3.1.2.) che equivale di fatto ad una tabella di un database in cui, per ogni *elemento* della rete stradale, è memorizzata una serie di campi relativi alla tipologia di tratto, quali veicoli vi possono circolare, le connessioni con gli altri tronchi, la disposizione dei numeri civici e tutte le informazioni necessarie a caratterizzare completamente il singolo elemento.

Analizzando i dati riportati nella mappa è stato possibile modellare la rete stradale con un grafo non orientato utile per le successive elaborazioni. I nodi sono associati agli incroci mentre gli archi ai tronchi stradali esistenti fra un incrocio ed il successivo.

Associata alla mappa, i cui dati rappresentano la struttura della rete di viabilità, sono riportate in un database separato le informazioni relative alle manovre proibite; queste sono indicate come coppie ordinate di elementi della rete stradale: la tupla $\langle a, b \rangle$ indica il divieto di passare dal tratto stradale a a quello b nel caso in cui questi due si intersechino presso un incrocio.

2.3 Procedure di elaborazione dati

L'obiettivo dell'analisi ed elaborazione dei dati spaziali e temporali è arrivare a definire lo scenario in cui il Servizio "118" si troverà ad operare. I modelli sviluppati nel Capitolo 3 possono essere studiati sul lungo periodo, sia settimanale, stagionale o annuale, a scelta del decisore. La singola esecuzione però, richiede di operare su parametri costanti che definiscano completamente le condizioni dello scenario, particolari condizioni di criticità o lo stato del traffico. Si tratta quindi di individuare delle frazioni del periodo di attività riportato dallo storico all'interno delle quali il sistema sia in condizioni di stazionarietà rispetto alle seguenti grandezze:

- λ^v, frequenza di generazione di una richiesta di soccorso non urgente (un codice verde);
- λ^u, frequenza di generazione delle richieste urgenti (codici gialli e rossi) che prevedono il soccorso entro il tempo limite di 8 minuti;
- $\overline{d}$, distanza massima, espressa in chilometri, chele ambulanze possono coprire entro la soglia degli 8 minuti;
- μ_v, frequenza di completamento delle missioni associate ad un codice verde;

- μ_u, frequenza di completamento delle missioni urgenti, associate a codici gialli e rossi.

Le procedure descritte in questo paragrafo sono state realizzate a questo scopo: elaborare i dati forniti dalla centrale operativa per calcolare delle frazioni del tempo di attività del servizio, dette *fasce temporali*, all'interno delle quali i parametri che descrivono lo scenario si possono ragionevolmente considerare costanti. Le informazioni estratte in questa fase sono fondamentali per lo studio dei modelli per il supporto alle decisioni, presentati nei capitoli 3 e 4.

Nei paragrafi 2.3.1 e 2.3.2 vengono descritti due strumenti realizzati per essere integrati nelle procedure di elaborazione presentate infine nel Paragrafo 2.3.3.

2.3.1 Un algoritmo per il geocoding

Come si è visto nel Paragrafo 2.1, nello storico di servizio sono riportati diversi campi che permettono di localizzare geograficamente un evento sul territorio. Nel registro relativo ai punti di stazionamento, il cui schema è mostrato nella Tabella 2.2, nei due campi VL_RIF_X e VL_RIF_Y sono riportate le coordinate geografiche della posizione della struttura. Grazie a queste informazioni è possibile localizzare i punti d'attesa sul territorio, così come visto per gli ospedali in Figura 2.2.1. Oltre che per le strutture, i campi contenenti le coordinate servono alla localizzazione degli eventi registrati nello storico di attività. Nei registri relativi agli interventi di soccorso (Tabella 2.4) ed alle tratte percorse dai mezzi (Tabella 2.3) sono presenti i campi VL_RIF_X e VL_RIF_Y che ne permettono il posizionamento. È quindi possibile sapere da quale punto del territorio è stata generata una data richiesta di soccorso o quali sono i luoghi di partenza ed arrivo di una tratta.

Con delle coordinate geografiche rilevate in modo preciso, posizionare un punto sulla mappa del territorio sarebbe un compito banale. Durante l'analisi delle informazioni geografiche riportate nello storico sono state rilevate però numerose incongruenze: indirizzi in aree disabitate, punti localizzati in zone senza diverse da dove si trova l'indirizzo associato, tratte dei mezzi in zone irraggiungibili. Inoltre, la presenza simultanea di tutte le informazioni di posizionamento (via, civico e coordinate) si ha solamente in poche migliaia di tuple; nella totalità dei casi c'è un'ambiguità di difficile soluzione riguardante i nomi delle vie; è così che Via Giacomo Leopardi può essere indicata con il nome completo così come Via G. Leopardi, Via Leopardi G. o solamente Via Leopardi. Questa scarsa affidabilità è probabilmente imputabile all'attuale sistema di immissione dati in dotazione alle centrale: l'operatore, rice-

vuta la chiamata di soccorso, seleziona a mano il punto sulla mappa dell'area di competenza. Qualsiasi scostamento, soprattutto se in zone in cui la rete stradale è molto densa, può portare ad un grave errore di localizzazione. Per questo motivo si è scelto di definire un algoritmo di *geocoding* robusto, che riuscisse a superare l'ambiguità data dalla scarsa accuratezza dei dati memorizzati in alcune tuple dello storico.

Tradizionalmente, un algoritmo di geocoding riceve in ingresso delle coordinate geografiche e restituisce come risultato l'attribuzione di un indirizzo (via, numero civico, ecc...) mentre un algoritmo di *reverse geocoding* procede in senso contrario, dato un indirizzo postale ritorna le presunte coordinate di quel punto geografico. Il metodo definito in questo lavoro è un algoritmo che si può definire come "ibrido" in quanto, sfruttando tutte le informazioni disponibili fra nome della via, numero civico e coordinate geografiche, tenta di posizionare in modo affidabile il punto sulla mappa per poi assegnarlo ad un elemento del grafo corrispondente ad un luogo della rete stradale.

Le informazioni che possono essere disponibili sono: le coordinate geografiche, il nome della via ed il numero civico. La prima indica un punto sulla mappa e, nonostante la localizzazione possa essere estremamente imprecisa, le coordinate sono presenti in tutte le tuple dello storico. La seconda informazione, il nome della via, non è sempre presente e, come già accennato, può essere indicata in innumerevoli formati ed abbreviazioni differenti. Il numero civico è l'informazione più rara da trovare nel registro degli eventi e, nel caso sia riportato, è presente sempre in coppia con il nome della via. L'algoritmo di *geocoding* sviluppato in questo lavoro opera con qualsiasi combinazione di dati.

Caso I

Nel primo caso l'algoritmo deve lavorare avendo a disposizione solamente le coordinate geografiche del punto. In questa situazione viene utilizzata la tecnica di geocoding normalmente adottata da tutti i sistemi GIS: il punto viene associato all'elemento a distanza minima. Il funzionamento di questa prima versione è descritto in Algoritmo 1 dove d_{min} rappresenta la distanza minima rilevata fra il punto dato p e l'elemento della rete stradale più vicino, indicato con e^*. Le due funzioni $\texttt{buffer}(p, \mathfrak{s})$ e $\texttt{distanza}(p, e)$ sono fornite entrambe dal GIS: la prima permette di estrarre dalla mappa tutti gli elementi della rete stradale che cadono, anche in minima parte, entro la distanza di soglia $\mathfrak{s}$ espressa in metri. Questo valore è fissato a seconda del livello di fiducia che si ripone nei dati da analizzare: più $\mathfrak{s}$ è piccola, più si richiederà al punto p di essere vicino ad un elemento della rete viaria. Se nessun elemento

Algoritmo 1 : geocoding versione 1, solo le coordinate geografiche a disposizione.

```
1:  procedure ASSEGNAMENTO1(p, 𝔰)
2:      d_min = ∞
3:      e* = ∅
4:      for each e ∈ buffer(p, 𝔰) do
5:          d = distanza(p, e)
6:          if d < d_min then
7:              d_min = d
8:              e* = e
9:          end if
10:     end for
11:     if e* ≠ ∅ then
12:         l = lato(p, e*)
13:     end if
14:     return {e*, l}
15: end procedure
```

risulta entro distanza di $\mathfrak{s}$ metri, p viene considerato impossibile da posizionare in modo sicuro. Se alla fine dell'esecuzione dell'algoritmo si ha che $e^* = \varnothing$, si può decidere di ripetere il geocoding aumentando la soglia o dichiarare il punto definitivamente non assegnabile.

La seconda funzione utilizzata, $\texttt{distanza}(p, e)$, calcola la distanza in metri che intercorre fra il punto p e l'elemento stradale e.

Se alla fine dell'esecuzione si ha che $e^* \neq \varnothing$, ovvero è stato trovato almeno un elemento entro la distanza di soglia, grazie alla funzione $\texttt{lato}(p, e^*)$ si ricava il lato del tratto stradale sul quale posizionare p. L'assegnamento avviene in modo semplice: se p si trova nel semipiano destro identificato dalla retta di cui e^* è segmento, allora il lato sarà il destro, altrimenti sarà il sinistro.

Caso II

Nel secondo caso i dati a disposizione sono le coordinate del punto e l'indirizzo incompleto, comprendente solo il nome della via in cui il punto dovrebbe essere posizionato. Questa variante, descritta in Algoritmo 2, è stata realizzata estendendo la prima versione, introducendo la capacità di valutare la somiglianza di due nomi. Questa modifica è stata necessaria in quanto, a causa della scarsa precisione del posizionamento dei punti degli eventi, può capitare che l'elemento più vicino non sia quello voluto e che questa ambiguità possa essere risolta confrontando i nomi. L'approccio scelto è stato quello di calcolare, per ogni elemento

Algoritmo 2 : geocoding versione 2 con coordinate geografiche ed indirizzo parziale (solo il nome della via) a disposizione.

```
 1:
 2: procedure ASSEGNAMENTO2(p, n_p, 𝔰, 𝔱)
 3:     d_min = ∞
 4:     t* = 0
 5:     e* = ∅
 6:     for each e ∈ buffer(p, 𝔰) do
 7:         n_e = nome(e)
 8:         if n_e ≠ ∅ then
 9:             d = (𝔰 − distanza(p, e)) /𝔰
10:             t = somiglianza(n_p, n_e) + d
11:             if t > t* ∧ t > 𝔱 then
12:                 t* = t
13:                 e* = e
14:             end if
15:         end if
16:     end for
17:     if e* ≠ ∅ then
18:         l = lato(p, e*)
19:     end if
20:     return {e*, l}
21: end procedure
22:
```

candidato, un “punteggio” di somiglianza al quale contribuiscono sia il grado di affinità del nome che la distanza dal punto.

Come si nota dalla descrizione dell’algoritmo, oltre alla soglia in metri che regola la distanza massima (s) è stato aggiunto un secondo valore limite, indicato con t, che indica il punteggio minimo che un elemento deve avere per essere considerato un candidato valido. Alla riga 9 (Algoritmo 2) viene calcolato il valore del punteggio di distanza in modo tale che, più è lontano l’elemento dal punto, più il valore finale viene penalizzato. Alla riga 10 viene calcolato il valore del punteggio finale come la somma fra il valore di “affinità” fra l’indirizzo parziale riportato nello storico (n_p) e quello dell’elemento corrente (n_e). Si noti che la funzione $\texttt{nome}(p)$ recupera il nome associato all’elemento che le viene passato come argomento mentre, di particolare importanza, è la funzione $\texttt{somiglianza}(n_p, n_e)$ che calcola il livello di affinità fra i due nomi, parti di indirizzi, esprimendolo con un valore reale fra 0 e 1. Ricavato il punteggio finale t associato all’elemento corrente e, se questo supera sia il miglior punteggio calcolato fino a quel momento (t^*) che la soglia imposta sull’affinità (t), viene considerato come il candidato più promettente.

Così come avviene per il Caso I, se è stato trovato almeno un elemento candidato valido si procede a determinare geometricamente il lato di appartenenza del punto.

Caso III

Nell’ultimo caso preso in esame, le informazioni a disposizione sono complete: nello storico sono presenti le coordinate del punto, il nome della via ed il numero civico. L’assegnamento del punto ad un elemento della rete stradale avviene come visto nell’Algoritmo 2; il dato aggiuntivo disponibile, il numero civico, permette di stabilire con maggiore sicurezza quale lato dei due sensi di marcia assegnare il punto.

Per ognuno degli elementi della mappa, il GIS conserva una serie di informazioni relative ai sensi di marcia, agli intervalli di numeri civici ed alla loro disposizione. In particolare, per le mappe utilizzate in questo lavoro i campi sono:

- `*_RefAddr`: il numero civico presente all’estremo iniziale del tratto stradale;
- `*_NRefAddr`: il numero civico presente all’estremo finale;
- `*_AddrSch`: lo schema di numerazione seguito dai civici. Il valore di questo campo può essere `E` per indicare che i civici su quel lato della strada hanno valori unicamente pari, `O` per indicare valori dispari mentre con `M` si avranno valori di qualsiasi tipo.

Si noti che ognuna delle voci è specificata separatamente per il lato destro e sinistro del tratto stradale, al simbolo * va sostituita la lettera R per il lato destro ed L per il sinistro.

Sfruttando queste informazioni è possibile decidere scegliere il lato in modo più preciso. Nel caso in cui il civico riportato nello storico non sia compatibile (ad esempio se il suo valore è pari ma gli schemi di entrambi i lato sono di tipo O) o, di contro, compatibile con entrambi i lati (ad esempio se gli schemi di indirizzamento destro e sinistro sono di tipo M ed i valori agli estremi comprendono il civico del punto) si scarta l'informazione aggiuntiva in quanto inutile e si ritorna ad utilizzare l'algoritmo visto per il Caso II.

Conclusione

All'interno dello storico relativo alle missioni dell'anno 2005 sono state registrate 247297

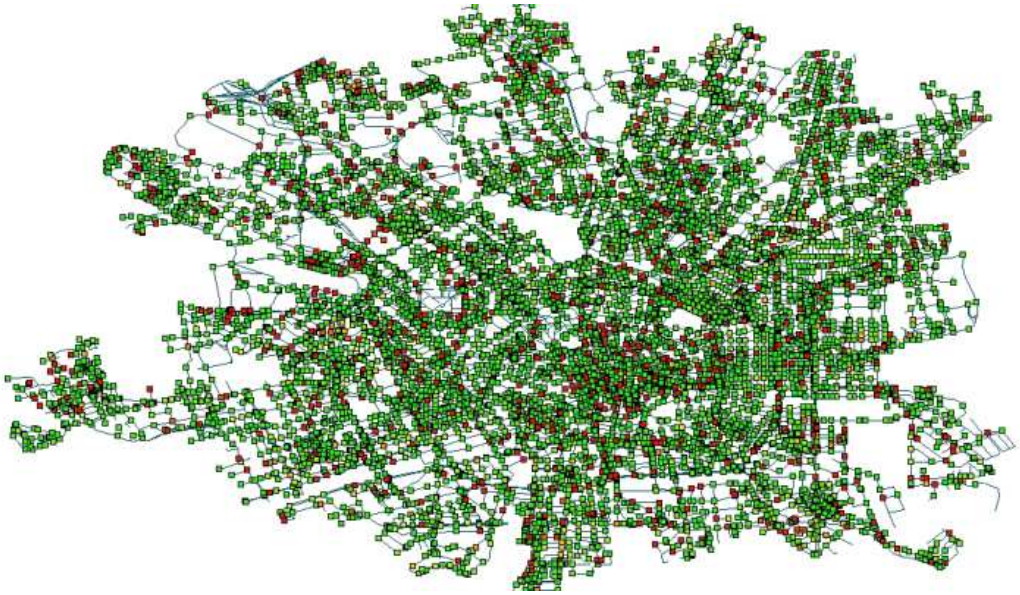


Figura 2.4: classificazione, effettuata tramite geocoding, dei nodi della città di Milano per frequenza di richieste.

missioni che hanno richiesto un intervento dell'ambulanza. Di queste, 323 sono associate a punti del tipo I solo con le coordinate geografiche, 54375 sono del tipo II con solo l'indirizzo ma senza il numero civico mentre i restanti 192599 appartengono al tipo III, hanno quindi tutti i dati a disposizione.

Grazie alla realizzazione di questo algoritmo è stato possibile integrare nelle procedure di elaborazione dati la localizzazione sul territorio di qualsiasi evento riportato nello storico

di attività. In Figura 2.4 ad esempio, è riportato il risultato della classificazione di tutti i nodi del grafo dell'area urbana di Milano a seconda della loro criticità: questo è stato possibile grazie all'utilizzo di questo strumento, il quale consente di associare qualsiasi chiamata di emergenza al relativo luogo sulla rete stradale. Il metodo di geocoding visto in questo paragrafo permette di fatto l'unione della dimensione temporale (gli eventi) a quella spaziale e geografica (la posizione).

2.3.2 Un algoritmo per la costruzione delle fasce temporali

L'obiettivo delle procedure sviluppate in questo capitolo è ottenere una suddivisione temporale in *fasce*, all'interno delle quali i parametri relativi al tasso di interarrivo delle richieste ed alla condizione del traffico siano ragionevolmente considerati costanti.

Un primo approccio al problema è stato quello di costituire delle fasce di lunghezza fissa pari a 30 minuti, sufficientemente piccola da garantire una ragionevole precisione ma, nel contempo, sufficientemente grande da contenere un numero ragionevole di eventi; i parametri sono stati quindi calcolati per ognuna di queste prime fasce di base. Si noti che una fascia può essere identificata come una tupla contenente i parametri che la definiscono completamente: $\langle \lambda^v,\, \lambda^u,\, \mu_v,\, \mu_u,\, \overline{d} \rangle$.

Successivamente si è voluto costruire fasce più grandi sulla base di quelle iniziali, aggregando fra loro quelle che potevano essere considerate compatibili al fine di suddividere la singola giornata in un numero di macro-fasce N dato, ognuna di dimensione variabile. A questo scopo è stato definito un algoritmo di programmazione dinamica che, operando sulla suddivisione basilare in 30 minuti, aggrega le fasce iniziali tentando di massimizzare la dimensione delle aree temporali risultanti ma, al contempo, minimizzando la varianza dei dati aggregati.

Si consideri la giornata suddivisa in n fasce iniziali; l'algoritmo di programmazione dinamica usa la ricorsione sulla seguente

$$t_i^*(t_{i+1}) = \min_{t_{i-1}+1 \le t_i \le t_{i+1}-1} \left\{ \sigma^{2^*}(t_i) \;+\; \sigma^2(t_i,\, t_{i+1}) \right\} \qquad \forall t_{i+1} = t_{i-1}+2, \ldots, t_n \quad (2.1)$$

per aggregare le fasce di base in N macro-suddivisioni. Con t_i si identifica la fascia di base che suddivide in modo ottimo le due adiacenti, la $t_{i-1},\, t_i$ e la $t_i,\, t_{i+1}$. L'algoritmo valuta tutte le possibili posizioni di t_i e fissa la suddivisione che minimizza la somma delle varianze delle due macro-fasce risultanti. Con σ^{2^*} però si identifica la varianza minima della fascia prece-

dente a t_i che per essere ricavata necessita di eseguire la 2.1 sull'intervallo $t_{i-1},\ t_i$. Ciò che succede è che quest'ultima viene fissata dall'esecuzione ricorsiva dell'algoritmo mentre al passo attuale è lasciato decidere come suddividere lo spazio non ancora partizionato, quello identificato da $t_i,\ t_{i+1}$. La profondità di ricorsione è data dal numero di macro-suddivisioni N, scelto per l'aggregazione delle fasce di base da 30 minuti. In questo modo, l'aggregazione delle fasce di base porta ad avere una suddivisione in N intervalli di dimensione arbitraria, la cui costruzione minimizza la varianza dei dati aggregati.

2.3.3 Le procedure realizzate

Utilizzando gli strumenti visti nei paragrafi precedenti è stato possibile realizzare delle procedure di elaborazione dati che permettono di ottenere la suddivisone in fasce temporali con caratteristiche ragionevolmente omogenee per quanto riguarda tutti i parametri.

Dal punto di vista della velocità di percorrenza, la costruzione fasce delle fasce ha richiesto:

1. l'estrazione dei dati riguardanti le tratte percorse dai mezzi, comprensive di tempi e coordinate geografiche dei punti di partenza e destinazione (si veda la Tabella 2.3);

2. il calcolo della velocità media di percorrenza di ognuna delle tratte. Per ricavare questo dato è stato necessario posizionare sul grafo i punti di partenza ed arrivo (tramite l'algoritmo di geocoding del visto nel Paragrafo 2.3.1), calcolare la distanza percorsa supponendo che fosse il cammino fosse quello a distanza minima (sfruttando gli algoritmi per il percorso stradale minimo) ed infine calcolare la velocità media avendo a disposizione tempo di percorrenza e distanza coperta;

3. la suddivisione preliminare in fasce statiche. Per facilitare l'elaborazione dei dati ricavati, ogni giornata è stata suddivisa in fasce di 30 minuti l'una, dimensione che, sia secondo l'opinione del decisore che sul campo, è risultata un giusto compromesso fra accuratezza e comodità;

4. l'aggregazione delle fasce di base in "macro-fasce" a varianza minima utilizzando il metodo di programmazione dinamica visto nel Paragrafo 2.3.2. L'utilizzo di questo modello ha permesso di massimizzare la dimensione delle fasce finali allo scopo di diminuire la mole di dati da trattare, nel contempo minimizzando la varianza dei valori di velocità media delle fasce di base (da 30 minuti) aggregate;

Per quanto riguarda la frequenza di occorrenza delle richieste di emergenza, la costruzione delle fasce ha previsto:

1. estrazione dal database storico dei dati riguardanti gli interventi di soccorso effettuati sul territorio, data ed ora di generazione e posizione geografica;

2. costruzione di un layer GIS da sovrapporre alla mappa stradale dell'area di competenza. In questo strato state riportate tutte le chiamate di soccorso localizzate sul territorio grazie all'algoritmo di geocoding visto nel Paragrafo 2.3.1;

3. suddivisione preliminare in fasce statiche. Come visto per i dati di velocità, per facilitare l'elaborazione dei dati ricavati ogni giornata è stata suddivisa in fasce di 30 minuti l'una;

4. aggregazione delle fasce di base in "macro-fasce" a varianza minima utilizzando il metodo di programmazione dinamica visto nel Paragrafo 2.3.2. L'utilizzo di questo modello ha permesso di massimizzare la dimensione delle fasce finali allo scopo di diminuire la mole di dati da trattare, nel contempo minimizzando la varianza dei valori di velocità media delle fasce di base (da 30 minuti) aggregate.

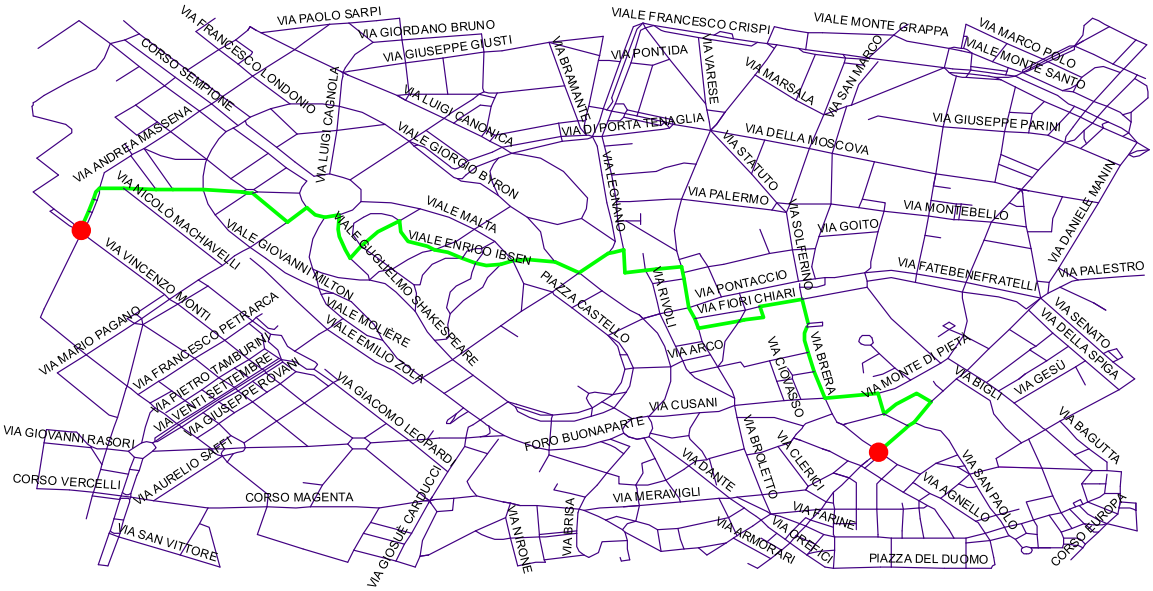


Figura 2.5: esempio di un percorso stradale minimo fra due punti di una mappa.

Si noti che, per calcolare il cammino minimo sulla rete stradale, è stato implementato un *algoritmo di Dijkstra bidirezionale*: questa variante utilizza un classico algoritmo di Dijkstra [8] che viene eseguito contemporaneamente in modo *forward*, etichettando cioè i nodi partendo dall'origine, sia in modo *backward*, etichettando i nodi partendo dalla destinazione. I due rami di ricerca procedono esattamente come visto in [8] ma la condizione di terminazione prevede che l'algoritmo si fermi quando si verifica la doppia etichettatura di un nodo. Il procedimento può essere descritto come:

1. esecuzione *forward* e *backward* dell'algoritmo di Dijkstra;

2. aggiornamento del valore dell'upper bound dato dal costo del percorso minimo trovato. All'atto di percorrere un arco che porta ad un nodo già etichettato dall'altra direzione, si può procedere solo se il percorso risultante migliora questo valore;

3. l'esecuzione viene fermata quando una delle due direzioni di ricerca etichetta un nodo già marcato dall'altra.

La particolarità della variante qui realizzata è stato il necessario rispetto da parte della ricerca del codice della strada. Il calcolo del cammino minimo stradale deve tenere conto delle manovre proibite, dei sensi di percorrenza e di tutte le altre limitazioni alla circolazione. Un esempio di percorso stradale è presentato in Figura 2.5; si noti che, nonostante nella raffigurazione origine e destinazione siano incroci, essi possono essere posizionati in qualsiasi punto lungo un arco.

Infine, per ottenere fasce ragionevolmente omogenee per tutti i parametri è stata operata un'intersezione fra i due insiemi di "macro-fasce", ottenendo così un insieme finale di suddivisioni associate a dati di velocità e domanda omogenei. In Figura 2.6 è rappresentato il diagramma di flusso delle attività messe in atto per implementare le procedure di elaborazione descritte.

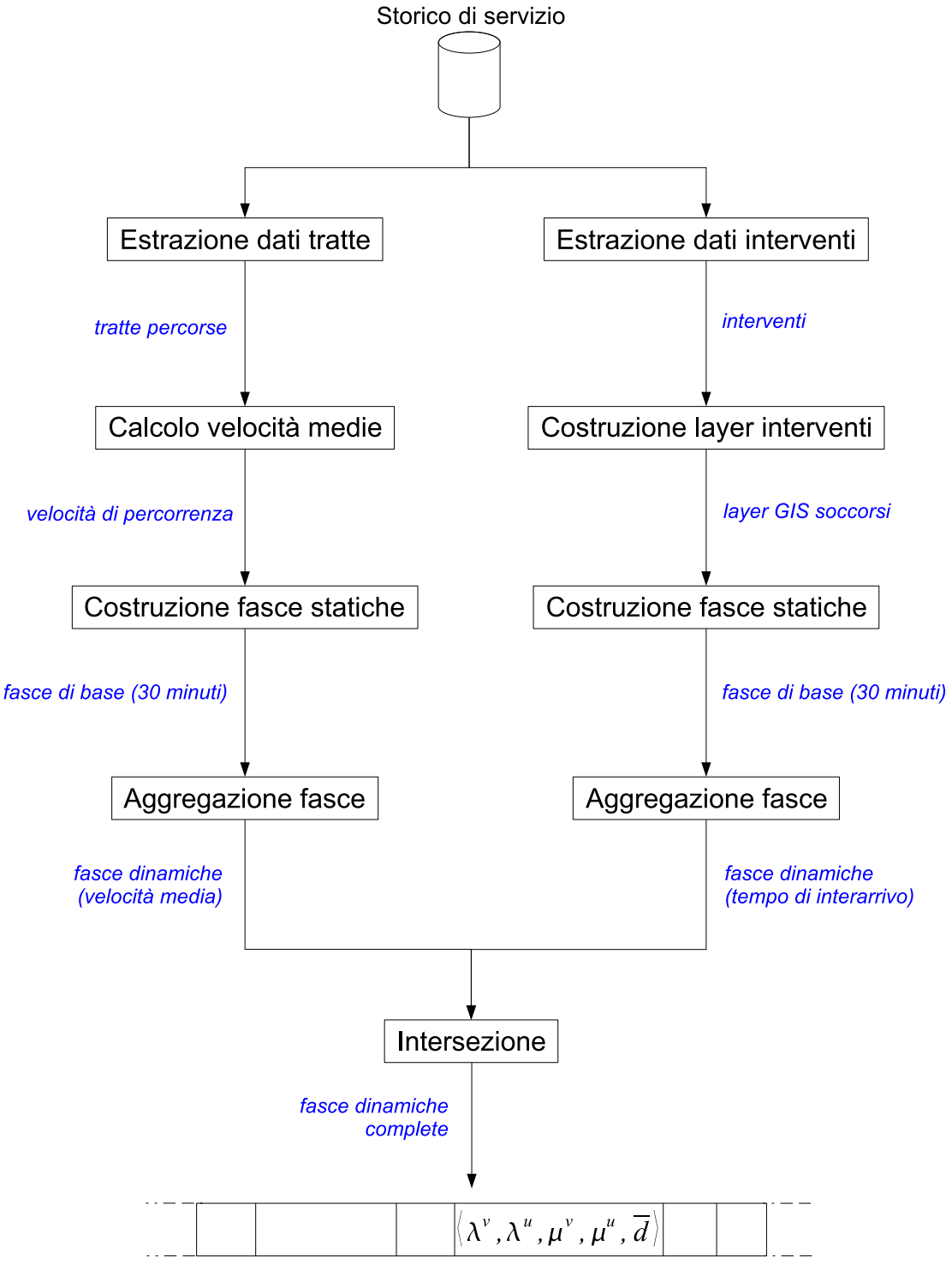


Figura 2.6: diagramma di flusso relativo alle procedure di elaborazione dati sviluppate.

Capitolo 3

Modelli per il supporto alle decisioni

Introduzione

Lo scopo di questa parte del lavoro di tesi è di fornire ai soggetti incaricati del dimensionamento del Servizio “118” uno strumento in grado di fungere da supporto durante la fase decisionale. Si è decisa la realizzazione di un insieme di modelli matematici il cui scopo è quello di permettere uno studio del comportamento del servizio al variare di alcune variabili decisionali ritenute significative:

- numero di mezzi di soccorso di proprietà dell’azienda ospedaliera. Questo valore, indicato in seguito con N, indica quante ambulanze sono disponibili senza richiedere il pagamento di alcun compenso; sono i mezzi dei quali si può disporre senza limitazioni di utilizzo.

- numero di mezzi di soccorso di terzi. É la quantità disponibile (indicata in seguito con R) di ambulanze di proprietà di terze parti, principalmente private, che richiedono il pagamento di un compenso per il noleggio della risorsa. La politica di tariffazione adottata prevede che sia possibile inviare gratuitamente il mezzo presso un punto d’attesa. Solo l’effettivo impiego dell’ambulanza (comprensiva di equipaggio) in una missione di soccorso richiede il pagamento di un compenso.

- numero di punti d’attesa (colonnine). É il numero di punti di stazionamento che il servizio è in grado di allestire. In seguito verrà indicato con C.

- livello di guardia per le chiamate urgenti. Indicato con K, rappresenta il numero di ambulanze libere, non impegnate in missione, al di sotto del quale il sistema serve solo chiamate urgenti (codici gialli e rossi). Se il numero di mezzi in attesa scende fino a K, le successive chiamate non urgenti (codici verdi) vengono accodate per mantenere sempre un insieme minimo di mezzi pronti a servire chiamate urgenti. Come spiegato in dettaglio nel Paragrafo 3.3, questo valore equivale anche alla lunghezza della coda.

Data la natura del servizio in esame, la scelta è caduta sull’utilizzo di modelli a coda: l’insieme di elementi di cui sono costituiti (analizzati in dettaglio nel Paragrafo 3.1) ha permesso una modellazione realistica ed adatta allo studio del comportamento del sistema. Un approccio simile a quello utilizzato in questo lavoro è quello di Larson ([19]) il quale, sfruttando le stesse basi modellistiche, realizza un modello *ad ipercubo* che permette il calcolo delle probabilità di occupazione dei mezzi lungo un periodo di tempo dato. Quanto definito

da Larson viene utilizzato prevalentemente per integrare elementi stocastici in modelli di programmazione matematica; a tal proposito si veda l'introduzione al Capitolo 4.

Con il termine *livello di servizio* si identifica la distribuzione delle probabilità fra i vari livelli di disponibilità dei mezzi di soccorso. In altre parole, per ogni possibile numero di ambulanze libere (ferme ad un punto d'attesa), è associato un valore di probabilità $x_{\langle n,t\rangle}$ con $n \in 0, \ldots, N$ numero di mezzi non occupati e $t \in 0, \ldots, T$ fascia oraria considerata fra le T complessive. Dati i valori per ogni $n \in 0, \ldots, N$, è possibile osservare con quale probabilità il servizio si troverà in ogni livello n in qualsiasi istante della fascia oraria t considerata. Un

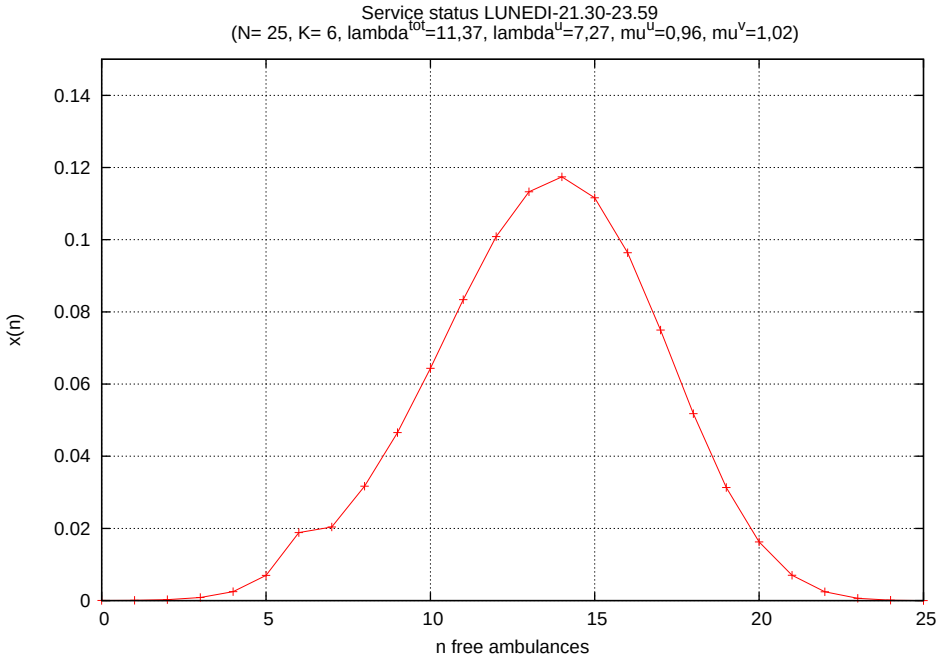


Figura 3.1: esempio di grafico del livello di servizio.

esempio di grafico del livello di servizio è mostrato in Figura 3.1. L'esempio è prodotto dal modello descritto in seguito nel Paragrafo 3.3: le ascisse rappresentano i livelli (discreti) dei mezzi liberi mentre sulle ordinate sono riportati i valori di probabilità $x_{\langle n,t\rangle}$. Lo scopo dei modelli per il supporto alle decisioni è quindi fornire una visione d'insieme del livello di servizio dati in ingresso valori arbitrari delle variabili decisionali d'interesse.

Nel seguito del paragrafo vengono presentati i modelli realizzati partendo da una prima versione (Paragrafo 3.2) che considera una sola variabile decisionale, per poi passare a considerare estensioni che aumentano il numero di variabili (Paragrafo 3.3 e Paragrafo 3.4) per

poi arrivare ad un modello completo e comprensivo di tutte le leve decisionali d'interesse (Paragrafo 3.5).

Come si vedrà nel seguito della trattazione, i modelli presentati considerano unicamente informazioni di tipo *temporale*, quello che producono è un vettore di valori di probabilità, equivalenti a frazioni del tempo complessivo della fascia oraria legate ad aspetti diversi; l'area di competenza è vista come un unico punto, non viene utilizzata alcuna informazione di tipo *spaziale*. Per questo motivo, i modelli trattati in seguito ricoprono un ruolo unicamente *strategico*, permettono al decisore di ampliare la sua capacità di visione sull'andamento del sistema, sulle sue prestazioni e possibili punti deboli, permettono di far fronte a situazioni straordinarie o critiche con un adeguato dimensionamento delle risorse. Per ampliare la capacità di supporto dei modelli, rendendoli utili anche a livello *tattico*, è necessario introdurre anche l'aspetto della consapevolezza spaziale. Considerare anche lo spazio comporterebbe comunque un enorme incremento della complessità; nonostante questo, lo sviluppo in questa direzione è prioritario negli obiettivi previsti per il progetto. Un primo passo è stato realizzare i modelli per l'ottimizzazione della copertura del territorio, trattati nel Capitolo 4: mettere in relazione i risultati di questi ultimi con quelli prodotti dai modelli a coda è la prospettiva più immediata di ampliamento.

3.1 Introduzione ai modelli a coda

La *Teoria delle Code* consiste nello studio, da un punto di vista matematico, di un fenomeno che si verifica in ogni situazione reale dove viene fornito un servizio e la domanda dei clienti può superare la capacità di erogazione. Come nel caso del Servizio "118", la predizione esatta di quando arriverà una richiesta è un obiettivo difficile da raggiungere. Per questo motivo, i processi decisionali che coinvolgono la capacità delle code d'attesa o il dimensionamento delle risorse risultano estremamente critici. Sovradimensionare il servizio può risultare troppo costoso in termini monetari (acquistare mezzi di soccorso che verranno utilizzati di rado, ricorrere a mezzi noleggiati più di quanto sia necessario). Di contro, sottodimensionare il servizio può essere molto pericoloso e portare a rischi per la salute dei cittadini. I sistemi a coda non risolvono direttamente i problemi della fase decisionale. Tuttavia, possono essere utilizzati con profitto per fornire informazioni preziose sulle prestazioni del servizio, diventando un utile strumento di supporto al decisore.

Elementi di un sistema a coda

Un sistema a coda è costituito da un insieme di servitori, deputati all'erogazione del servizio, ed un insieme di code d'attesa dove sono messi i clienti che non possono essere serviti immediatamente. Ogni cliente viene gestito da un singolo servitore. I processi che regolano

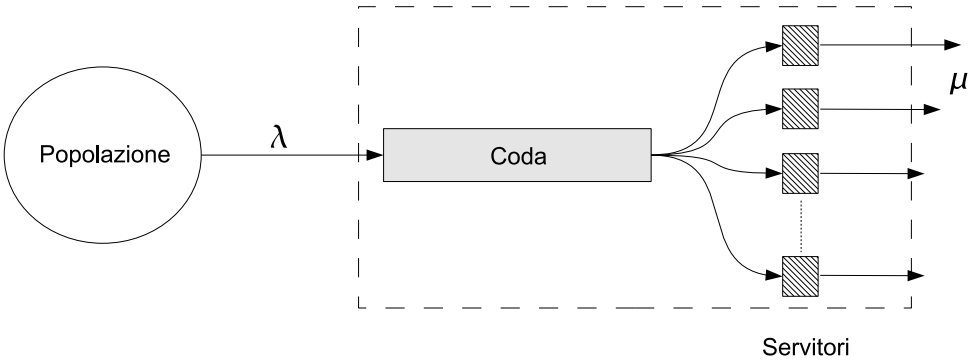


Figura 3.2: schema degli elementi di un generico sistema a coda.

il funzionamento del sistema a coda, mostrati schematicamente in Figura 3.2, sono:

- Popolazione: è l'insieme dei potenziali clienti che richiedono il servizio. Dato che generalmente i clienti di una popolazione sono omogenei, se esistono tipologie differenti si introducono popolazioni distinte. Dato che nel sistema preso in esame il processo di arrivo non è influenzato dal numero di clienti presenti nel sistema, la popolazione è da considerarsi *infinita*. Questo assunto risulta ragionevole: non è possibile sapere, in un dato istante, quante chiamate di soccorso arriveranno in seguito.

- Processo d'arrivo: è il processo stocastico che descrive come i clienti si presentano al sistema. É descritto dalla variabile aleatoria *tempo d'interarrivo*, ovvero il tempo intercorso fra due arrivi consecutivi. Come spesso si usa nel caso di sistemi a coda, anche in questo caso per garantire la trattabilità dei modelli si è assunto che il processo sia stazionario, ovvero che le sue proprietà statistiche non varino nel tempo. Per questo motivo è stata operata la suddivisione in fasce temporali del processo d'arrivo (Capitolo 2).

- Coda: è formata dai clienti che vengono posti in attesa data l'impossibilità di servirli al momento dell'arrivo.

- Servitori: come per i mezzi di soccorso, il loro numero è finito e costante. Dato che è fissato in fase di dimensionamento, il numero dei servitori è un'importante variabile decisionale. Ogni servitore è in grado di gestire un cliente per volta e, nel momento in cui il servizio è stato erogato, torna ad essere disponibile.

- Processo di servizio: descrive come i servitori soddisfano le richieste. É descritto dalla variabile aleatoria *tempo di servizio* che, come per il tempo di interarrivo, si considera ragionevolmente stazionaria.

- Disciplina di servizio: è la politica che specifica qual'è il primo cliente da servire fra quelli in attesa in coda. Fra le discipline presenti in letteratura ([16], p.66), quella che specifica meglio il Servizio "118" sembrerebbe essere la politica basata su classi di priorità: i codici assegnati ai vari casi (rosso, giallo e verde) sono di fatto valori di priorità. In realtà, come si vedrò per i modelli descritti in seguito, la coda contiene richieste omogenee. In questo caso, le richieste non urgenti (verdi) sono le uniche a poter essere messe in attesa senza pericolo per la persona. Dato che la coda contiene clienti a priorità omogenea, la disciplina di servizio è la *fifo* (first in, first out): le richieste vengono servite secondo l'ordine di arrivo.

Parametri di un sistema a coda

Le grandezze che caratterizzano un sistema a coda sono:

- *tasso di arrivo* λ, il valore atteso di arrivi per unità di tempo;

- *tasso di servizio* μ, il valore atteso dei completamenti per unità di tempo.

Si noti che in realtà λ è formalmente dipendente da n, numero di clienti presenti nel sistema, indicato con λ_n. Quando però λ_n è costante per ogni n, si ricade su λ. La stessa considerazione vale per μ: μ_n rappresenta il tasso di servizio per ogni servente occupato. Se è costante per ognuno diventa possibile utilizzare μ, come spiegato da Hillier e Lieberman ([16], p. 666). Date queste considerazioni, si ha che λ è pari all'inverso del tempo medio di interarrivo e μ equivale all'inverso del tempo medio di servizio.

Altre variabili sono significative per la descrizione del funzionamento di un sistema a coda:

- W è una variabile aleatoria che rappresenta il tempo medio di permanenza dei clienti nel sistema;
- L è il numero medio di clienti nel sistema;
- W_q è il valore medio del tempo di attesa in coda per ogni utente;
- L_q è il valore atteso della lunghezza della coda.

Queste, essendo variabili dipendenti dal tempo e dallo stato iniziale del sistema, risultano di difficile trattabilità analitica. È però dimostrato da Little in [20] che il sistema a coda, dopo una fase transitoria, si stabilizza ed i valori medi di L e W convergono a valori costanti indipendenti dal tempo. Il sistema converge all'equilibrio solo se

$$\rho = \frac{\lambda}{s\mu} < 1 \tag{3.1}$$

dove ρ è detto *fattore di utilizzo* ed s è il numero totale di servitori. Se ciò è verificato, si ha che

$$L = \lambda W$$

ovvero è valida la *formula di Little* che lega i valori di equilibrio (costanti) di L e W al parametro λ.

Il processo di nascita e morte

I modelli a coda realizzati in questo lavoro assumono che arrivo ed uscita dei clienti funzionino secondo un *processo di nascita e morte*: la *nascita* identifica l'arrivo di un cliente, la *morte* rappresenta l'uscita dal sistema una volta servito. In un processo di questo tipo, i due eventi avvengono casualmente, seguendo i valori medi di λ e μ ([16], p. 675). Un esempio di processo è riportato in Figura 3.3: gli archi rappresentano le possibili *transizioni* mentre i valori delle etichette sono il *tasso* medio di occorrenza di ogni transizione. L'indice n di ogni stato indica il numero di clienti nel sistema.

La condizione per il raggiungimento dell'equilibrio, espressa dall'equazione (3.1), continua ad essere necessaria. Se risulta soddisfatta, è verificato il principio chiave ([16], p. 667) secondo il quale *per ogni stato n, il tasso di uscita è uguale al tasso d'entrata*. Questo principio è rappresentato dall'*equazione di bilanciamento* per il generico stato n; considerando

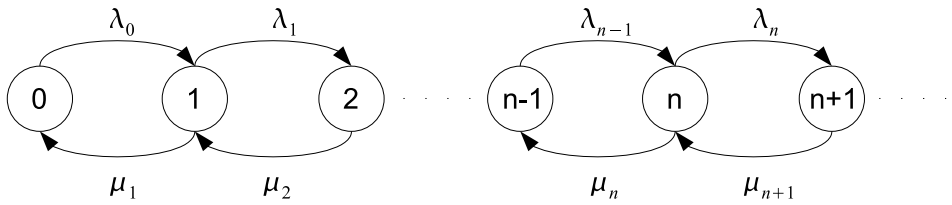


Figura 3.3: esempio di diagramma di transizione per un processo di nascita e morte.

infiniti stati possibili si ha che, all'equilibrio, la frequenza media di entrata in uno stato equivale a quella di uscita. In particolare, $x_{\langle n,t\rangle}$ *rappresenta la porzione del tempo (sul totale della fascia oraria* t*) in cui è possibile per il sistema trovarsi nello stato* n. Per $n \geq 1$ si ha

$$\mu x_{\langle n+1,t\rangle} + \lambda x_{\langle n-1,t\rangle} = (\mu + \lambda) x_{\langle n,t\rangle} \tag{3.2}$$

dove $x_{\langle n,t\rangle}$ indica la probabilità associata allo stato $\langle n, t\rangle$, riferito alla fascia temporale t in cui si hanno n clienti nel sistema. Per lo stato iniziale, $n = 0$, si ha:

$$\mu x_{\langle 1,t\rangle} = \lambda x_{\langle 0,t\rangle} \tag{3.3}$$

quindi, la probabilità di passare dagli stati $n + 1$ ed $n - 1$ è uguale a quella della transizione opposta. Si noti che le equazioni (3.2) e (3.3) possono essere espresse usando λ e μ costanti in quanto, come già provato da Larson in [18], i modelli realizzati in questo lavoro fanno parte della classe *M/M/s* (processo d'arrivo *Poissoniano*, processo di servizio esponenziale, s numero noto e finito di serventi). In [16] si prova che, per questa classe di processi di nascita e morte, il tasso medio d'arrivo λ_n ed il tasso medio di servizio per servitore occupato μ_n sono *costanti indipendentemente dallo stato del sistema*. Un esempio è mostrato in Figura 3.4.

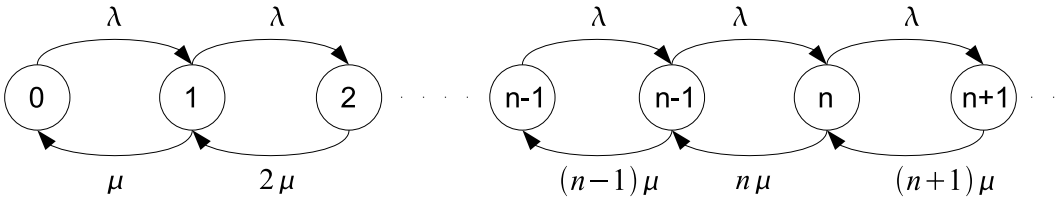


Figura 3.4: esempio di diagramma di transizione per un processo di classe *M/M/s*.

Dalle etichette degli archi di transizione si evince che, se il tasso medio di servizio per ogni

servitore occupato è μ, il tasso medio di servizio complessivo per n servitori occupati è $n\mu$. Anche in questo caso, le considerazioni fatte valgono finché l'equazione (3.1) è soddisfatta.

Nell'ambito dei processi di classe *M/M/s* i parametri visti nel Paragrafo 3.1 vengono calcolati in modo differente:

$$L_q = \frac{x_{\langle N \rangle} \left(\lambda/\mu\right)^N \rho}{N! \left(1-\rho\right)^N} \tag{3.4}$$

$$W_q = \frac{L_q}{\lambda} \tag{3.5}$$

$$W = W_q \frac{1}{\mu} \tag{3.6}$$

$$L = L_q \frac{\lambda}{\mu} \tag{3.7}$$

Note sulla distribuzione esponenziale

L'ipotesi più comune, al contempo semplice da trattare e realistica, per modellare le variabili tempo di interarrivo e di servizio è che la loro distribuzione sia di tipo esponenziale. Oltre a godere di utili proprietà ([16] p. 668-674), essa è immune da effetti di aggregazione e disaggregazione dei processi d'arrivo:

$$\prod_{i=1}^{k} e^{-\lambda_i t} = e^{-\sum_{i=1}^{k} \lambda_i t} \tag{3.8}$$

Questa proprietà ha permesso, nell'ambito dei modelli mostrati nei Paragrafo 3.3 e 3.5 di scomporre λ in più frazioni distinte per gravità dell'urgenza.

3.2 Dimensionamento della flotta dei mezzi di soccorso

Il primo approccio al Servizio "118" è consistito nella realizzazione di un modello a coda che permettesse la valutazione del livello di servizio esistente. In questo primo caso, l'unica variabile decisionale considerata è il numero di mezzi di soccorso disponibili N.

Il modello utilizzato è un semplice processo di nascita e morte a più serventi, come descritto nel Paragrafo 3.1. Dato N come numero totale di mezzi, si ha che $s = N - n$, ovvero: ogni stato ha indice n, numero di mezzi *disponibili*; di conseguenza il numero di

servitori impegnati nello stato $\langle n \rangle$ è pari a $N - n$. I tassi di transizione per $0 < n < N$ sono:

$$in(n) = x_{\langle n-1 \rangle}(N - (n-1))\mu + x_{\langle n+1 \rangle}\lambda$$
$$out(n) = x_{\langle n \rangle}(\lambda + (N-n)\mu)$$

con $out(n)$ tasso complessivo delle transizioni in uscita dallo stato n, $in(n)$ tasso complessivo in entrata. L'equazione di bilanciamento risulta quindi:

$$in(n) = out(n)$$

da cui si ha

$$x_{\langle n-1 \rangle}(N - (n-1))\mu + x_{\langle n+1 \rangle}\lambda = x_{\langle n \rangle}(\lambda + (N-n)\mu) \tag{3.9}$$

Un esempio di modello è riportato in Figura 3.5. Dal diagramma si nota come l'equa-

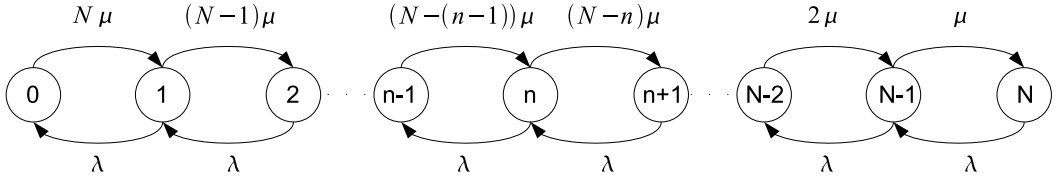


Figura 3.5: esempio di diagramma di transizione per un modello della prima versione.

zione (3.9) sia valida solo per $0 < n < N$. Gli stati particolari si hanno per $n = N$ ed $n = 0$:

$$x_{\langle n-1 \rangle}\mu - x_{\langle n \rangle}\lambda = 0 \qquad \text{con } n = N \tag{3.10}$$
$$x_{\langle n+1 \rangle}\lambda - n\, x_{\langle n \rangle}\mu = 0 \qquad \text{con } n = 0. \tag{3.11}$$

Si noti che agli estremi è possibile cancellare gli archi delle transizioni inesistenti in quanto, in quegli stati di frontiera, i valori di μ e λ per gli archi da eliminare hanno valore nullo.

A questo punto è possibile impostare un sistema lineare nelle $x_{\langle n \rangle}$ variabili di probabilità combinando le equazioni (3.9), (3.10) e (3.11) ed aggiungendo un vincolo che normalizzi i valori delle $x_{\langle n \rangle}$ probabilità:

$$\sum_{n=0}^{N} x_{\langle n \rangle} = 1 \tag{3.12}$$

Il sistema risultante è lineare e facilmente risolubile con l'ausilio di un solutore di programmazione lineare.

3.2.1 Risultati

Questo primo modello permette di analizzare il livello di servizio raggiunto dal sistema al variare del numero di ambulanze messe in campo. Il sistema è stato scritto in linguaggio *GNU MathProg* (si veda [15]) e risolto con l'utilizzo del solutore lineare *GLPK* (*GNU Linear Programming Kit*). Dati una fascia oraria t ed il valore di N, il modello produce in uscita il grafico del livello di servizio che visualizza i valori delle probabilità $x_{\langle n \rangle}$ per ogni $n \in 0 \dots N$. In Figura 3.6 è riportata una serie di grafici del livello di servizio tutti riguardanti la stessa fascia oraria ($\lambda = 14,440$, $\mu = 0,830$) ma per ognuna delle situazioni è stata usata una differente dimensione della flotta di soccorso. Come si può notare, all'aumentare dei mezzi disponibili, il valore massimo di $x_{\langle n \rangle}$ si sposta verso N. Questo comportamento è ragionevole: aumentando i mezzi impiegati e lasciando λ costante si ottiene la diminuzione della probabilità di avere poche ambulanze libere. Il decisore può procedere imponendo sia un valore massimo sul numero di ambulanze occupate (non si vuole che il sistema passi troppo tempo estremamente carico), sia un valore minimo (non si vuole nemmeno sovradimensionare la flotta) e scegliere un valore di N che concentri i valori massimi di $x_{\langle n \rangle}$ fra i due limiti. In Figura 3.7 sono riportati i grafici prodotti dal modello nel caso in cui venga lasciata N costante e si faccia variare la fascia oraria. Come si può vedere, al crescere di λ (la fascia diventa sempre più critica) i valori massimi di $x_{\langle n \rangle}$ si spostano verso lo zero. Anche in questo caso il comportamento è sensato: all'aumentare del tasso d'arrivo si ha l'aumento delle frazioni di tempo in cui ci sono più ambulanze impegnate in missione. Il decisore potrebbe operare in questo modo per valutare la capacità di una flotta di far fronte a situazioni sempre più critiche. Si noti che, come spiegato nel Paragrafo 3.1, ha senso la risoluzione del sistema solo per valori di N, λ e μ che soddisfano l'equazione (3.1), la condizione di equilibrio (dove N è indicato con s).

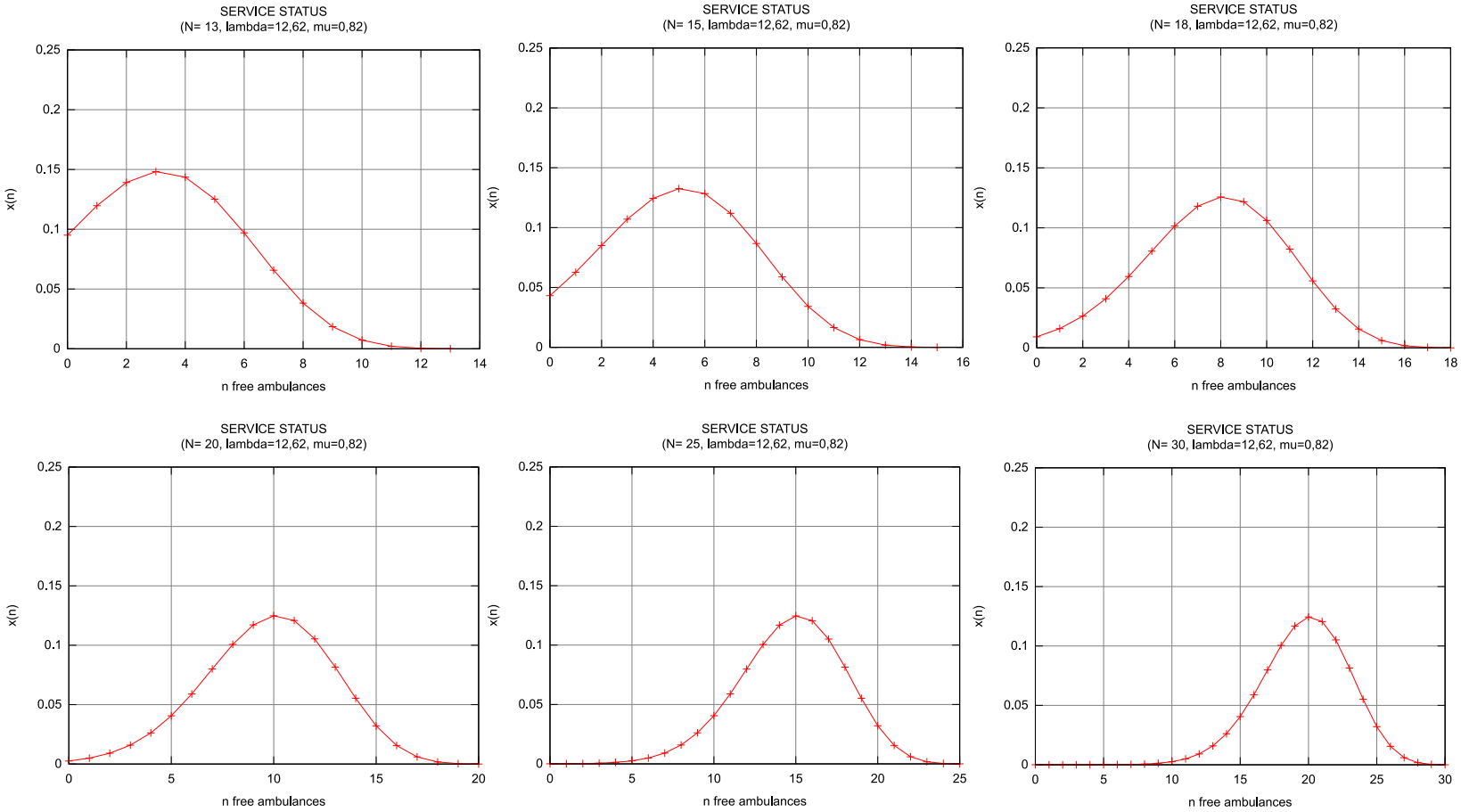


Figura 3.6: esempio di grafici del livello di servizio prodotti al variare del parametro N.

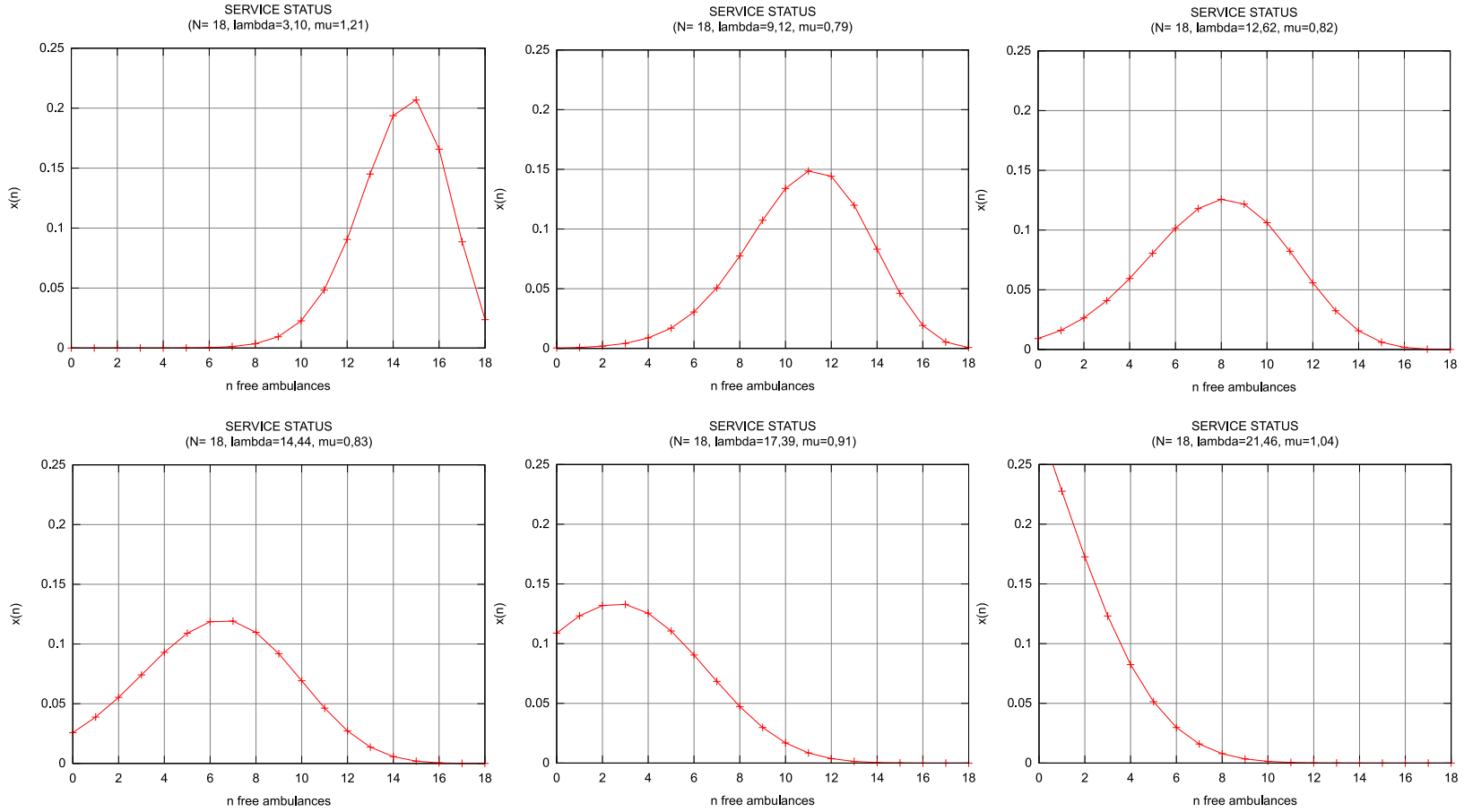


Figura 3.7: esempio di grafici del livello di servizio prodotti al variare del parametro λ.

3.3 Valutazione della politica di gestione delle chiamate non urgenti

Il secondo approccio al problema prevede l'utilizzo del modello visto nel paragrafo precedente, da estendere per introdurre una nuova leva decisionale. Questa seconda versione ammette esplicitamente la possibilità per il servizio di accodare le chiamate identificate come codici verdi. Queste ultime sono, come introdotto nel Capitolo 1, le uniche a non prevedere pericolo di vita per la persona e, di conseguenza, le uniche a non essere soggette a vincoli di tempestività del servizio. Oltre a fornire controllo sulla dimensione della flotta, è stata aggiunta una nuova variabile che, indicata con K, rappresenta il limite minimo di ambulanze libere al di sotto del quale il servizio entra in *zona critica*. In ogni stato dove $n \leq K$ è consentito accodare le richieste con codice verde per dedicare tutti i mezzi liberi alla gestione delle richieste urgenti (codici gialli e rossi).

3.3.1 Disaggregazione dei processi di nascita e morte

Negli stati che si trovano in zona critica, al sistema è consentito accodare le richieste "verdi" mentre le chiamate urgenti devono essere servite immediatamente. Questo comportamento richiede la suddivisione dei processi d'arrivo e di servizio per calcolare correttamente i tassi di transizione differenziati.

Per la proprietà espressa dall'equazione (3.8), è possibile separare il processo d'arrivo:

$$\lambda^{tot} = \lambda^u + \lambda^v \tag{3.13}$$

dove λ^{tot} è la frequenza d'arrivo complessiva, λ^u è la frequenza d'arrivo di richieste urgenti (codici gialli e rossi) e λ^v la frequenza d'arrivo delle richieste non urgenti (codici verdi).

Lo stesso criterio è stato adottato per la disaggregazione del processo di servizio:

$$\mu^{tot} = \left(\frac{\lambda^u}{\lambda^{tot}}\right)\mu^u + \left(\frac{\lambda^v}{\lambda^{tot}}\right)\mu^v \tag{3.14}$$

dove μ^u è il tasso di servizio delle richieste urgenti, μ^v rappresenta il tasso di servizio delle "verdi". Entrambi i valori sono pesati con la frazione di richieste dello stesso tipo rispetto al totale per ottenere il tasso di servizio complessivo μ^{tot}; in particolare, i termini λ^u/λ^{tot} e

λ^v/λ^{tot} sono il numero medio di ambulanze occupate a servire richieste associate ai codici dei due tipi (urgenti per il primo, verdi per il secondo).

3.3.2 Il modello

A differenza di quanto realizzato per il modello precedente, le dimensioni di indicizzazione degli stati sono due: $n \in 0, \ldots, N$ per i mezzi liberi, $k \in 0, \ldots, K$ per le richieste "verdi" messe in attesa. In Figura 3.8 è rappresentato schematicamente un modello con $K = 3$ ed

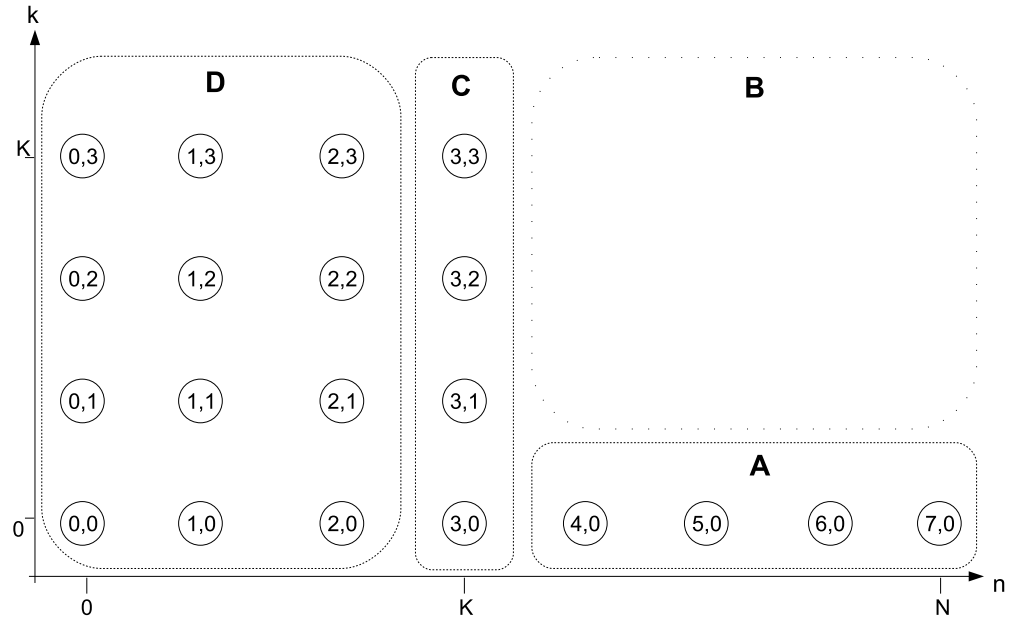


Figura 3.8: struttura e suddivisione in aree degli stati del modello a coda non urgente.

$N = 7$. Gli stati possono essere suddivisi in insiemi a seconda delle caratteristiche delle equazioni di bilanciamento.

$\mathbf{A} \equiv \{\, \langle n, k \rangle \;:\; n \in 0, \ldots, N;\, k \in 0, \ldots, K;\, \mathbf{n} > \mathbf{K};\, \mathbf{k} = \mathbf{0} \,\}$
Per gli stati appartenenti all'insieme A valgono le considerazioni fatte per il modello di Paragrafo 3.2: le equazioni (3.10) e (3.11) ne descrivono completamente i tassi di transizione. Si noti che è $k = 0$ costante in quanto, non esistendo in A stati in zona critica, non è permesso al

sistema l'accodamento di richieste "verdi". In questo caso, dato che non è necessario distinguere le tipologie di richieste, nei tassi di transizione vengono usati i processi indifferenziati λ^{tot} e μ^{tot}.

$\mathbf{B} \equiv \{\, \langle n,k \rangle \;:\; n \in 0,\ldots,N;\; k \in 0,\ldots,K;\; \mathbf{n} > \mathbf{K};\; \mathbf{k} > \mathbf{0} \,\}$
Gli stati presenti in questo insieme non sono ammissibili: non è possibile che vengano accodate richieste "verdi" fuori dalla zona critica. Dato che in B il numero di ambulanze libere è n con $n > K$, in questo insieme non si è mai autorizzati all'accodamento. Per rendere "inaccessibili" questi stati viene introdotto un vincolo aggiuntivo

$$x_{\langle n,k \rangle} = 0 \qquad \forall n \in 0,\ldots,N;\; k \in 0,\ldots,K \;:\; n > K,\; k > 0 \tag{3.15}$$

che forza a zero la probabilità associata ad ogni stato appartenente all'insieme B.

$\mathbf{C} \equiv \{\, \langle n,k \rangle \;:\; n \in 0,\ldots,N;\; k \in 0,\ldots,K;\; \mathbf{n} = \mathbf{K} \,\}$
Gli stati appartenenti all'insieme C rappresentano una situazione di transizione: il sistema è già entrato in zona critica ma, nel caso si liberi un mezzo, questo deve essere usato immediatamente per servire una richiesta precedentemente accodata. In Figura 3.9 è rappresentata una generica configurazione per l'insieme C. Come si vede, lo stato $\langle K,0 \rangle$ è l'unico in cui, a fronte della liberazione di un mezzo, non si ha un immediato riutilizzo; nelle altre due tipologie, la $\langle K,K \rangle$ e la generica $\langle K,k \rangle$, la disponibilità di un'ambulanza provoca una transizione verso il basso, ovvero il servizio di una richiesta "verde" precedentemente accodata. In $\langle K,0 \rangle$ un mezzo che si libera non ha nulla da fare perché la coda è vuota ($k = 0$): l'ambulanza ritorna ad essere disponibile per future emergenze (transizione verso destra, rientro nell'insieme A). In tutti gli stati di questo insieme si è però in zona critica: la nascita di un cliente urgente provoca l'occupazione di un mezzo (transizione verso sinistra), l'arrivo di una chiamata non urgente fa scattare una transizione verso l'alto, aumentando di uno il numero dei clienti accodati. Le equazioni di bilanciamento risultano:

$$\begin{aligned}
&x_{\langle n+1,k \rangle} \lambda^{tot} + \\
&x_{\langle n,k+1 \rangle} \left(N-K\right) \mu^{tot} + \\
&x_{\langle n-1,k \rangle} \left(\left(N-K\right) \mu^{tot} + \mu^{u} \right) - \\
&x_{\langle n,k \rangle} \left(\left(N-K\right) \mu^{tot} + \lambda^{v} + \lambda^{u} \right) = 0 \qquad \text{con } n = K, k = 0
\end{aligned}$$

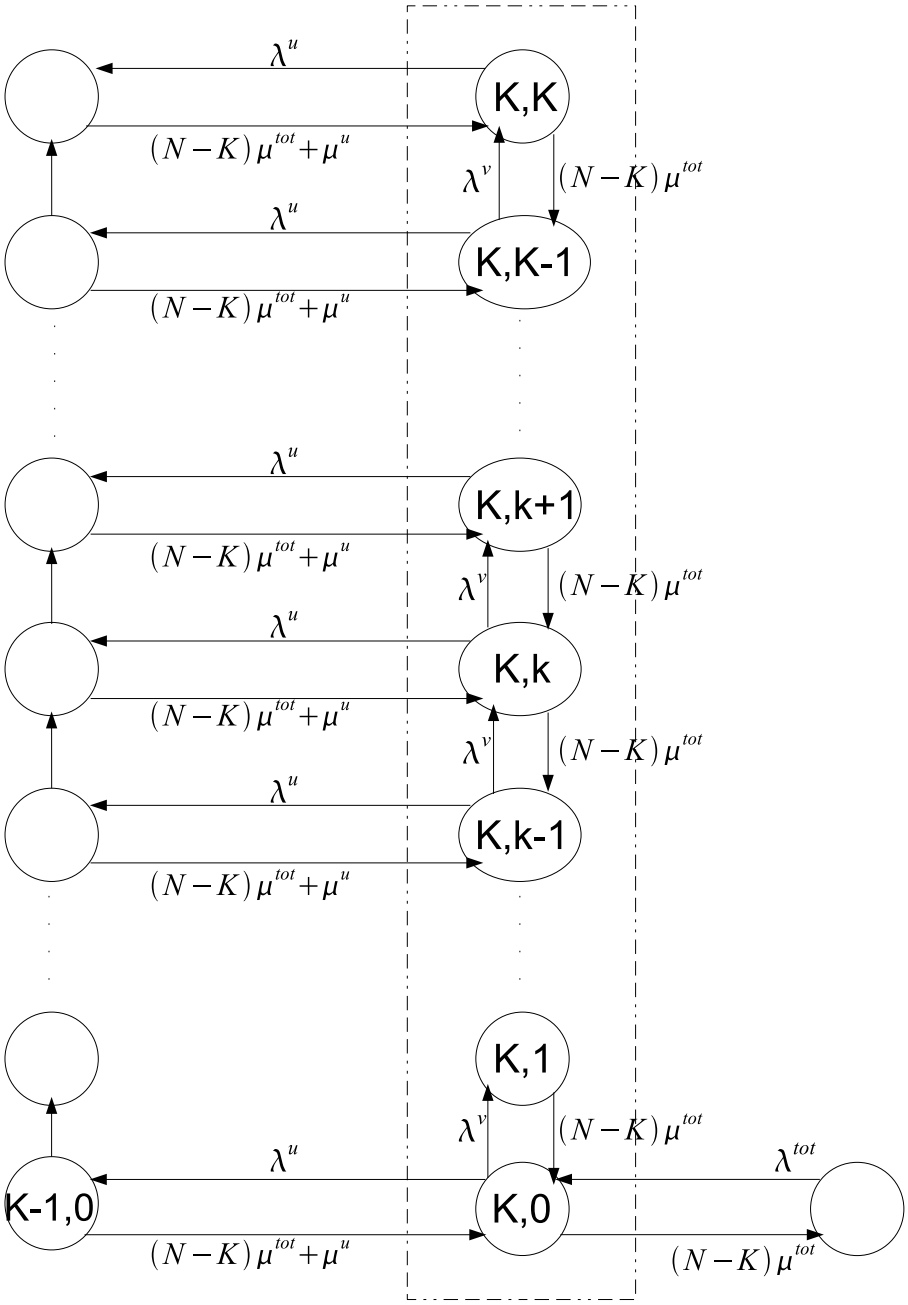


Figura 3.9: struttura e transizioni degli stati dell'insieme C.

$$
\begin{aligned}
& x_{\langle n,k-1\rangle}\lambda^v + \\
& x_{\langle n,k+1\rangle}\,(N-K)\,\mu^{tot} + \\
& x_{\langle n-1,k\rangle}\left((N-K)\,\mu^{tot}+\mu^u\right) - \\
& x_{\langle n,k\rangle}\left((N-K)\,\mu^{tot}+\lambda^v+\lambda^u\right) = 0 \qquad \text{con } n = K, 0 < k < K
\end{aligned}
$$

$$
\begin{aligned}
& x_{\langle n,k-1\rangle}\lambda^v + \\
& x_{\langle n-1,k\rangle}\left((N-K)\,\mu^{tot}+\mu^u\right) + \\
& x_{\langle n,k\rangle}\left((N-K)\,\mu^{tot}+\lambda^u\right) = 0 \qquad \text{con } n = K, k = K
\end{aligned}
$$

con $n \in 0,\ldots,N$ indice del numero di mezzi liberi e $k \in 0,\ldots,K$ indice della lunghezza della coda di richieste "verdi".

$\mathbf{D} \equiv \{\, \langle n,k\rangle \;:\; n \in 0,\ldots,N;\; k \in 0,\ldots,K;\; \mathbf{n} < \mathbf{K} \,\}$
Gli stati appartenenti al gruppo D sono tutti in zona critica. In questa situazione il sistema può allocare un'ambulanza solo per servire richieste urgenti; l'arrivo di una chiamata classificata con codice verde provoca sempre un accodamento. Come si vede in Figura 3.10, non

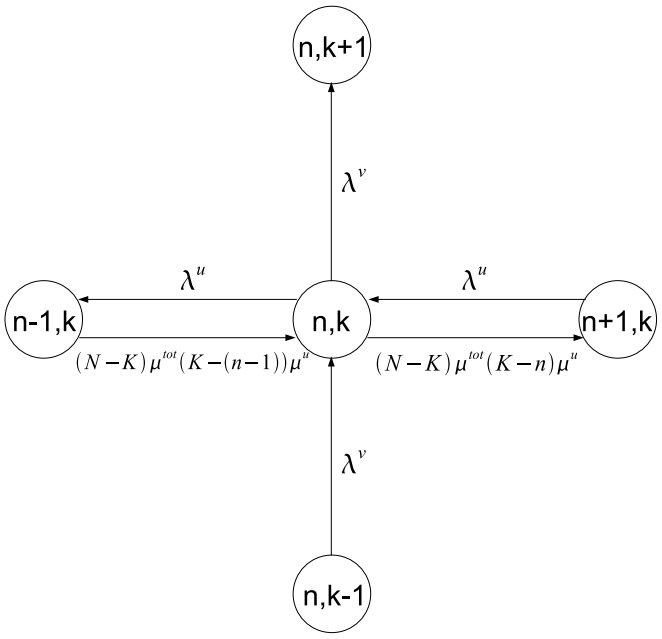


Figura 3.10: grafo delle transizioni di un generico stato dell'insieme D.

è possibile servire richieste non urgenti: anche nel caso in cui si liberi un mezzo, questo non

provoca lo svuotamento della coda (come invece avviene per gli stati dell'insieme C). Negli stati appartenenti all'insieme D l'ambulanza che ritorna libera viene lasciata disponibile per assicurare un servizio tempestivo alle chiamate "gialle" e "rosse": l'arco di transizione verso il basso è assente. L'unico modo per iniziare a servire le utenze rimaste in coda è liberare sufficienti mezzi per entrare negli stati del gruppo C. Le equazioni di bilanciamento, differenziate per gli stati di bordo, sono:

$$
\begin{aligned}
&x_{\langle n+1,k\rangle}\lambda^u + \\
&x_{\langle n-1,k\rangle}\left((N-K)\,\mu^{tot} + (K-n+1)\,\mu^u\right) - \\
&x_{\langle n,k\rangle}\left(\lambda^v + \lambda^u + (N-K)\,\mu^{tot} + (K-n)\,\mu^u\right) = 0 \qquad \text{con } 0 < n < K, k = 0
\end{aligned}
$$

$$
\begin{aligned}
&x_{\langle n,k-1\rangle}\lambda^v + x_{\langle n+1,k\rangle}\lambda^u + \\
&x_{\langle n-1,k\rangle}\left((N-K)\,\mu^{tot} + (K-n+1)\,\mu^u\right) - \\
&x_{\langle n,k\rangle}\left(\lambda^v + \lambda^u + (N-K)\,\mu^{tot} + (K-n)\,\mu^u\right) = 0 \qquad \text{con } 0 < n < K, 0 < k < K
\end{aligned}
$$

$$
\begin{aligned}
&x_{\langle n,k-1\rangle}\lambda^v + x_{\langle n+1,k\rangle}\lambda^u + \\
&x_{\langle n-1,k\rangle}\left((N-K)\,\mu^{tot} + (K-n+1)\,\mu^u\right) - \\
&x_{\langle n,k\rangle}\left(\lambda^u + (N-K)\,\mu^{tot} + (K-n)\,\mu^u\right) = 0 \qquad \text{con } 0 < n < K, k = K
\end{aligned}
$$

$$
\begin{aligned}
&x_{\langle n+1,k\rangle}\lambda^u - \\
&x_{\langle n,k\rangle}\left(\lambda^v + (N-K)\,\mu^{tot} + K\mu^u\right) = 0 \qquad \text{con } n = 0, k = 0
\end{aligned}
$$

$$
\begin{aligned}
&x_{\langle n,k-1\rangle}\lambda^v + x_{\langle n+1,k\rangle}\lambda^u - \\
&x_{\langle n,k\rangle}\left(\lambda^v + (N-K)\,\mu^{tot} + K\mu^u\right) = 0 \qquad \text{con } n = 0, 0 < k < K
\end{aligned}
$$

$$
\begin{aligned}
&x_{\langle n,k-1\rangle}\lambda^v + x_{\langle n+1,k\rangle}\lambda^u - \\
&x_{\langle n,k\rangle}\left((N-K)\,\mu^{tot} + K\mu^u\right) = 0 \qquad \text{con } n = 0, k = K
\end{aligned}
$$

con $n \in 0, \ldots, N$ e $k \in 0, \ldots, K$. Perché il modello sia completo è necessario aggiungere il vincolo di normalizzazione dei valori di probabilità $x_{\langle n,k\rangle}$ (come visto nell'equazione (3.12) a pagina 38).

3.3.3 Risultati

Anche in questo caso il sistema è stato scritto in linguaggio *GNU MathProg* e risolto con l'utilizzo del solutore *GLPK*. Oltre a fornire l'andamento del livello di servizio, questo modello ricava anche un'analoga distribuzione riguardante le frazioni di tempo nelle quali si ha la coda riempita ad ogni livello. Dato che i valori di $x_{\langle n,k \rangle}$ sono suddivisi su due dimensioni, per ricavare il livello di servizio è necessario sommare i valori sulle colonne:

$$x_{\langle n \rangle} = \sum_{k=0}^{K} x_{\langle n,k \rangle} \qquad \forall n \in 0, \ldots, N$$

Per ottenere invece i valori delle probabilità di avere k chiamate "verdi" in coda è necessario sommare i valori per riga:

$$x_{\langle k \rangle} = \sum_{n=0}^{N} x_{\langle n,k \rangle} \qquad \forall k \in 0, \ldots, K$$

Osservando l'output del modello, il decisore può valutare l'andamento delle prestazioni del servizio al variare del numero di mezzi e mantenendo il carico costante (Figura 3.11) e variando la soglia per l'accodamento delle richieste "verdi" (Figura 3.12); può osservare il comportamento della coda e del livello di servizio al variare della fascia oraria considerata (Figura 3.5). Oltre a valutare il livello di servizio (grafico di sinistra in ogni figura) il decisore può osservare per quali frazioni di tempo si hanno delle richieste che non vengono servite immediatamente (grafico di destra in ogni figura) e valutare l'accettabilità della loro incidenza. Dall'andamento dei grafici d'esempio si nota come, a fronte di una incrementata capacità di servizio, la coda tenda ad essere usata per una frazione sempre più ridotta del tempo. Una volta stabilita la dimensione della flotta, il decisore potrebbe tarare il livello di guardia per ottenere un utilizzo accettabile della coda per le chiamate non urgenti.

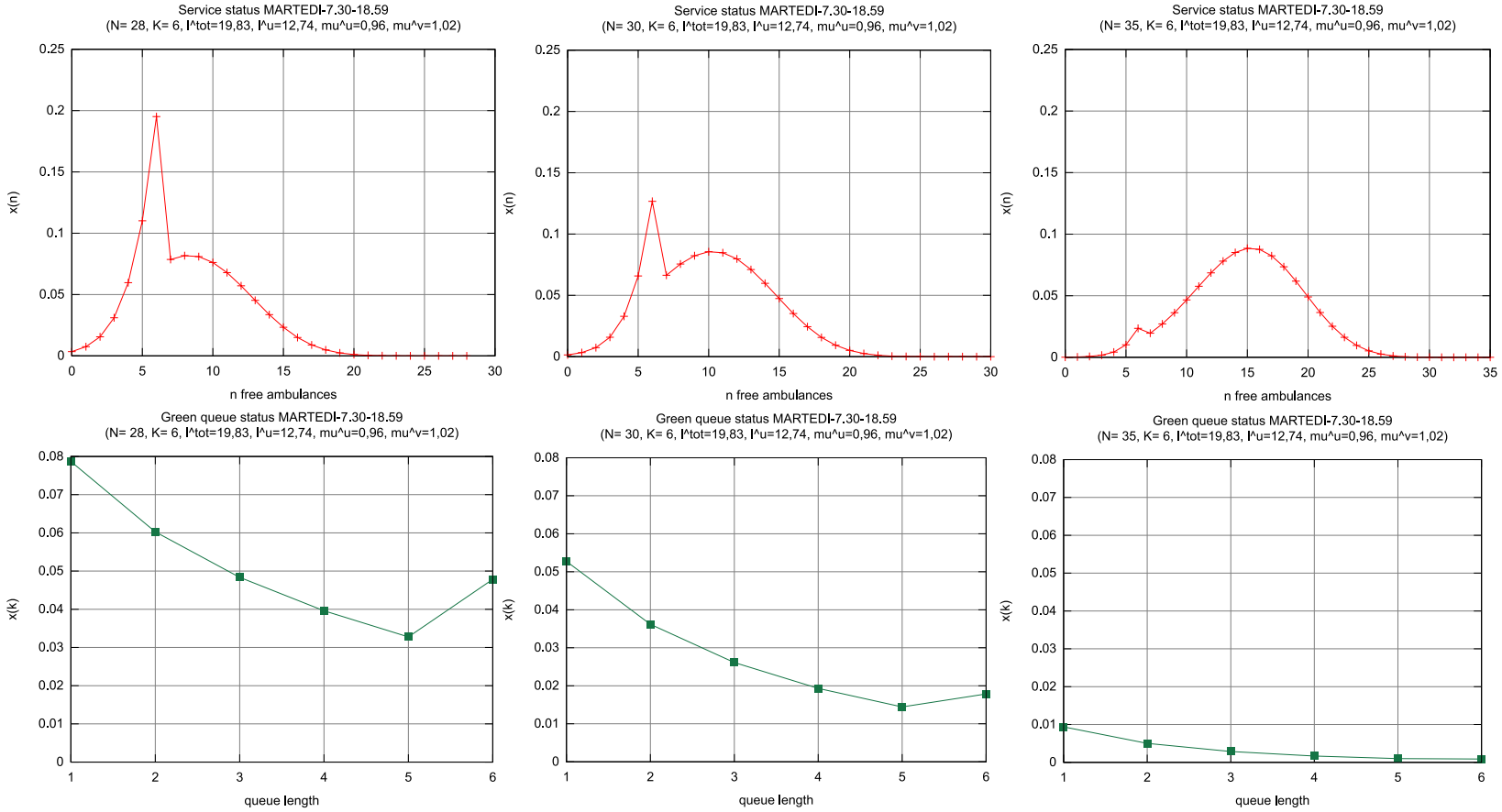


Figura 3.11: grafici del livello di servizio e stato coda al variare della dimensione della flotta.

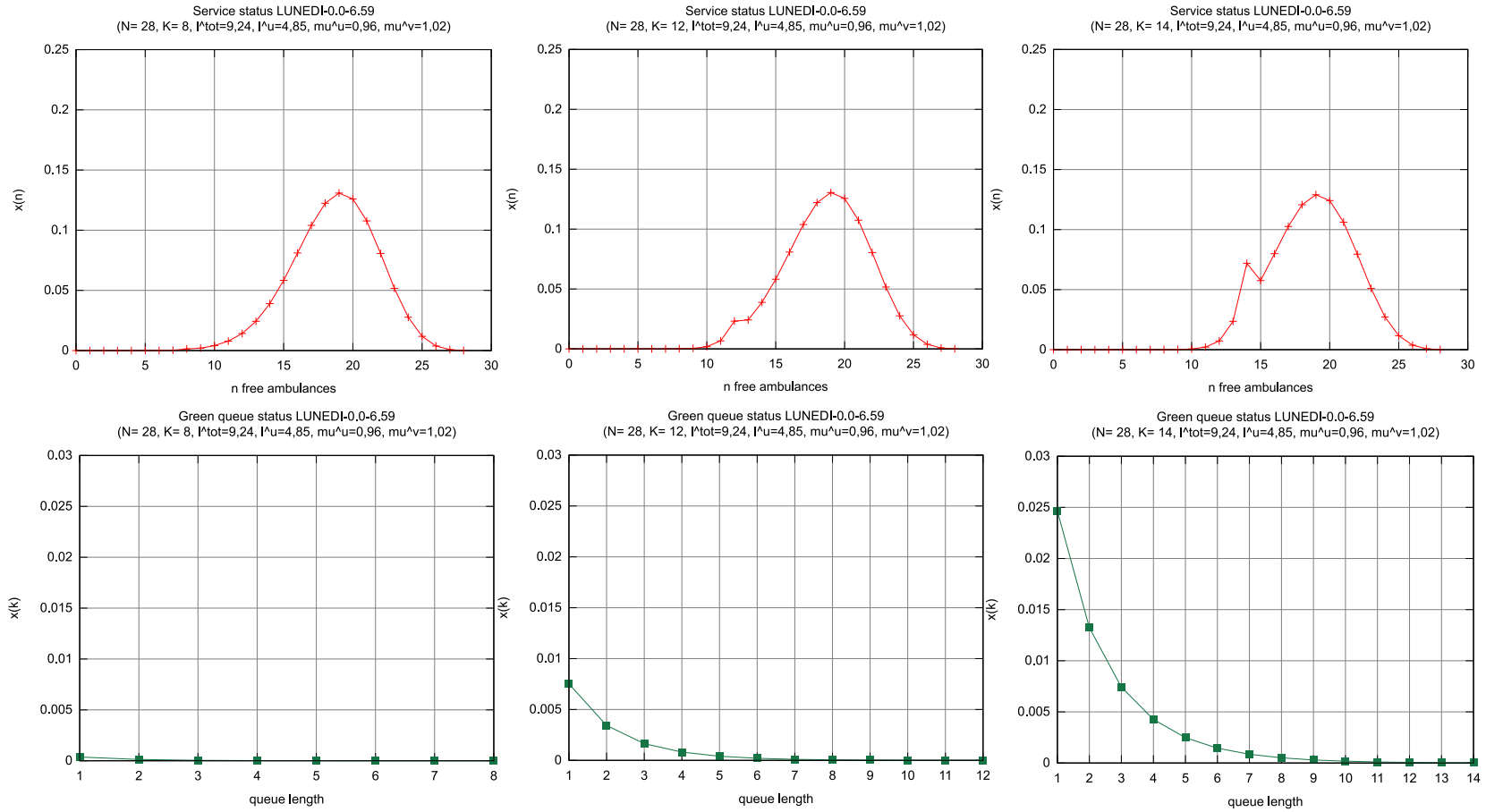


Figura 3.12: grafici del livello di servizio e stato coda al variare della soglia per la zona critica.

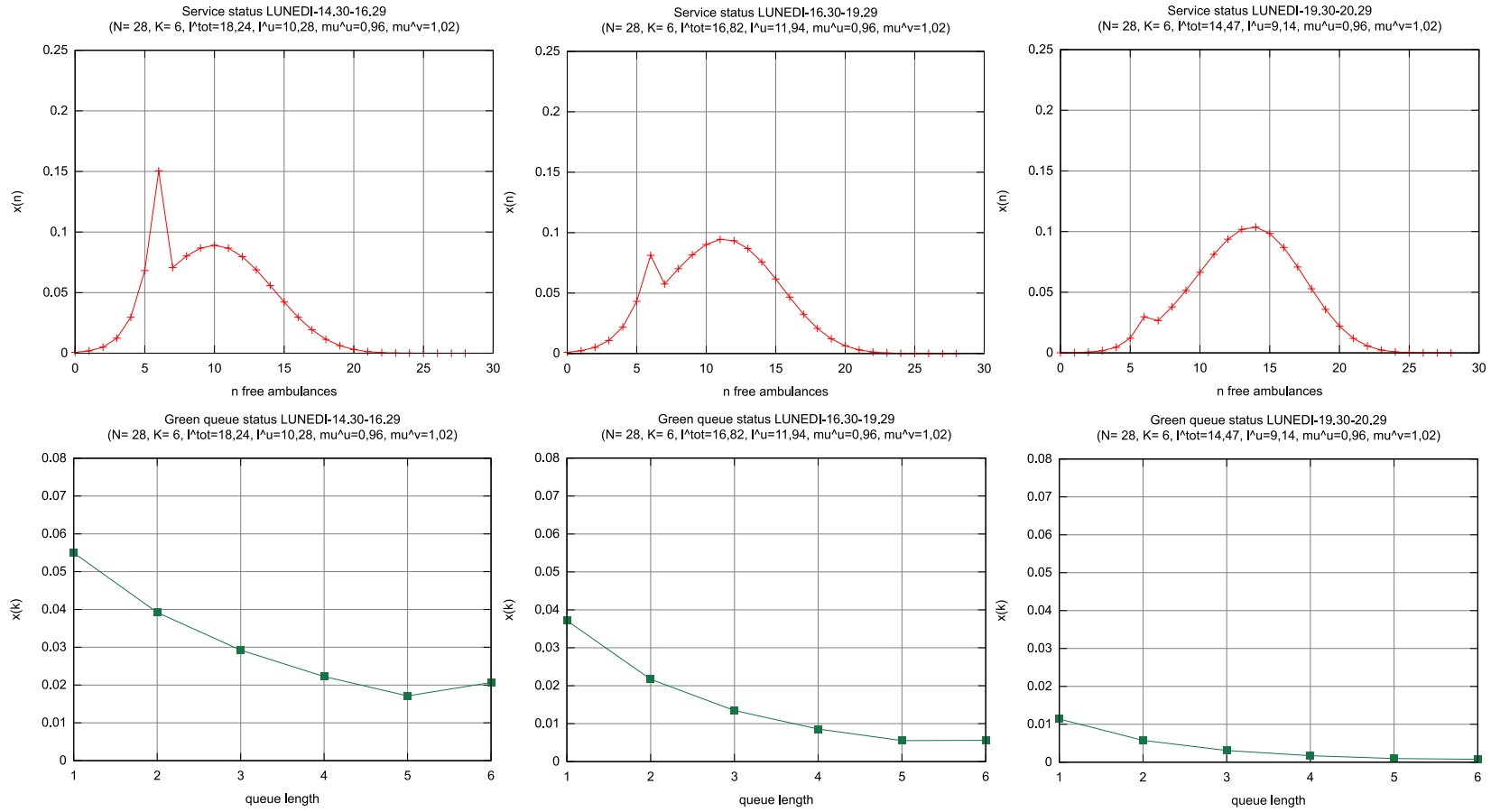


Figura 3.13: grafici del livello di servizio e stato coda al variare della fascia oraria considerata (variazione di λ).

3.4 Valutazione della politica di gestione dei mezzi di soccorso di terzi

Il terzo approccio al problema prevede l'estensione del modello visto nel Paragrafo 3.2 per aggiungere la possibilità di gestire i mezzi di proprietà di terze parti. Come introdotto nel Capitolo 1, questi sono mezzi di soccorso di associazioni private che vengono messi a disposizione dell'azienda sanitaria. Il loro utilizzo è vincolato al pagamento di un "gettone" di servizio, da versare solo a fronte di un effettivo impiego in missione dell'ambulanza e del suo equipaggio.

In questa variante le variabili decisionali coinvolte sono: la dimensione della flotta di ambulanze di proprietà dell'azienda sanitaria (indicata con N) e la dimensione della flotta di mezzi di proprietà di terzi (indicata con R). Inoltre è stato necessario introdurre un parametro che rappresentasse il numero di zone del territorio da servire (indicato con C) dove ognuna è coperta da un singolo mezzo. Il valore di C è usato come soglia di criticità in modo analogo a quanto avviene per il modello di Paragrafo 3.3: se il numero di mezzi liberi scende al di sotto di questo valore, il sistema è considerato in *zona critica* in quanto non è più in grado di coprire tutte le aree di competenza con ambulanze proprie. In questa situazione è consentito al Servizio "118" ricorrere all'utilizzo di mezzi di terze parti, le ambulanze dette "a gettone", per sopperire alla mancata copertura del territorio.

3.4.1 Nota sui tassi di transizione in zona critica

Dal momento che tutti i modelli trattati in questo capitolo non considerano informazioni spaziali, negli stati appartenenti alla zona critica, dove è possibile scegliere con quale tipologia di mezzo servire una richiesta, è necessario formulare un criterio che modelli in modo ragionevole la scelta. Si suppone quindi che la probabilità che una richiesta sia assegnata ad un'ambulanza di un dato tipo sia proporzionale alla quantità di mezzi di quel tipo rispetto al totale. Si ottiene quindi un tasso di transizione di nascita pari a:

$$\alpha_{n,r} = \lambda \frac{n}{n+r} + \lambda \frac{r}{n+r} \tag{3.16}$$

dove il primo termine pesa il tasso di interarrivo con il numero di mezzi di proprietà n normalizzato sul totale, il secondo termine pesa considerando il numero di ambulanze "a gettone" r. È importante notare che, sommando i due termini, si ottiene il tasso di interarrivo λ con

coefficiente pari a uno. Questo è sensato in quanto l'obiettivo dei due termini di pesatura è disaggregare le frequenze d'arrivo, nonostante il valore di λ sia indifferenziato per tipologia di servizio; in questo modo si suddivide il valore di frequenza in due componenti: la prima è la frequenza d'arrivo delle richieste che vengono servite da mezzi di proprietà, la seconda la frequenza d'arrivo delle richieste che vengono servite con un mezzo a noleggio.

Si consideri il fattore di normalizzazione usato nell'equazione (3.16): $n + r$ fornisce il totale delle ambulanze disponibili. Se però questo totale supera il numero di zone da coprire ($n+r \geq C$), i mezzi liberi sono sufficienti a servire tutto il territorio. Ciò porta a considerare come "ininfluente" il surplus di mezzi:

$$\alpha_{n,r} = \lambda\frac{n}{C} + \lambda\frac{C-n}{C} \tag{3.17}$$

dove si nota che la transizione associata al primo termine riguarda il servizio da parte di un mezzo proprio, il secondo riguarda il servizio a noleggio. Ciò è sensato: considerando uno stato in cui il numero di mezzi propri eguaglia il numero di zone ($n = C$), è certo che una nuova richiesta sarà sicuramente servita da una di queste ($n/C = 1$). Questo avviene indipendentemente da quanti mezzi di terze parti sono disponibili. Anche per questo caso valgono le considerazioni fatte per la (3.17): la somma dei coefficienti risulta pari a uno ma il loro scopo è quello di suddividere il valore di λ in due componenti, la prima relativa al servizio da parte dei mezzi di proprietà, la seconda relativa al servizio da parte di ambulanze "a gettone".

Alla luce di queste considerazioni, le equazioni (3.16) e (3.17) contribuiscono alla definizione del tasso per le transizioni di nascita

$$\alpha_n(n,r) = \begin{cases} \lambda\dfrac{n}{C} & \text{se } n+r \geq C, \\ \lambda\dfrac{n}{n+r} & \text{altrimenti} \end{cases} \tag{3.18}$$

$$\alpha_r(n,r) = \begin{cases} \lambda\dfrac{C-n}{C} & \text{se } n+r \geq C, \\ \lambda\dfrac{r}{n+r} & \text{altrimenti} \end{cases} \tag{3.19}$$

come mostrato in Figura 3.14: i tassi di transizione sono stati separati in due funzioni $\alpha_n(n,r)$ ed $\alpha_r(n,r)$, rispettivamente tasso di transizione di nascita lungo l'asse n (ambulanza di proprietà) e lungo l'asse r (mezzo a noleggio).

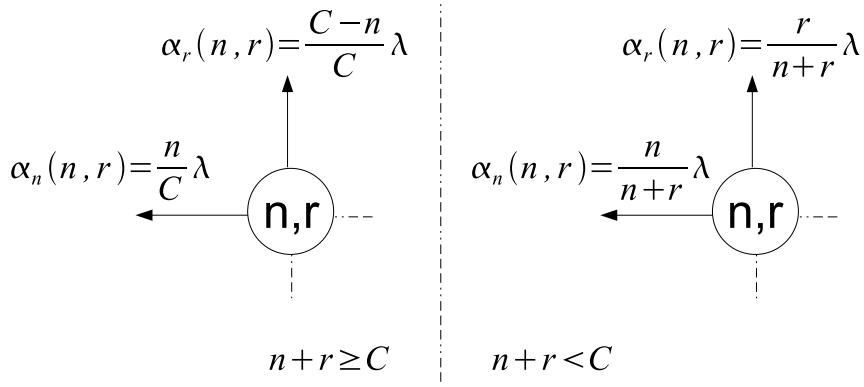


Figura 3.14: schema delle transizioni di nascita per gli stati in zona critica.

Per quanto riguarda le transizioni di servizio, queste vengono pesate sul numero di mezzi di ogni tipologia occupati nello stato:

$$o_{n,r} = (N - n)\,\mu + (R - r)\,\mu \tag{3.20}$$

dove il primo tasso riguarda le ambulanze di proprietà, il secondo quelle a noleggio. Con $o_{n,r}$ si identifica il tasso complessivo delle transizioni in uscita dallo stato $\langle n, r\rangle$.

3.4.2 Il modello

Le dimensioni di indicizzazione degli stati sono due: $n \in 0, \ldots, N$ numero di ambulanze di proprietà disponibili, $r \in 0, \ldots, R$ numero di mezzi "a gettone" non impiegati. In ognuno degli stati, $n + r$ dà quindi il totale dei mezzi a disposizione. In Figura 3.15 è rappresentata schematicamente la struttura degli stati in un modello dove $N = 8$, $R = 3$ e $C = 5$. Come si nota dallo schema, solo quando il sistema si trova in zona critica, dove i mezzi di proprietà non sono più sufficienti, è consentito al servizio il noleggio di ambulanze aggiuntive. Nel caso in cui ne venga usata comunque una di proprietà, avviene una transizione verso sinistra lungo l'asse n con un conseguente decremento del numero di mezzi propri. Se viene invece richiesto l'intervento di una terza parte, scatta una transizione verso l'alto, lungo l'asse r, facendo diminuire la quantità delle ambulanze "a gettone" disponibili.

Analogamente a quanto detto per il modello precedente, anche in questo caso gli stati possono essere raggruppati in insiemi distinti dalle diverse caratteristiche delle equazioni di

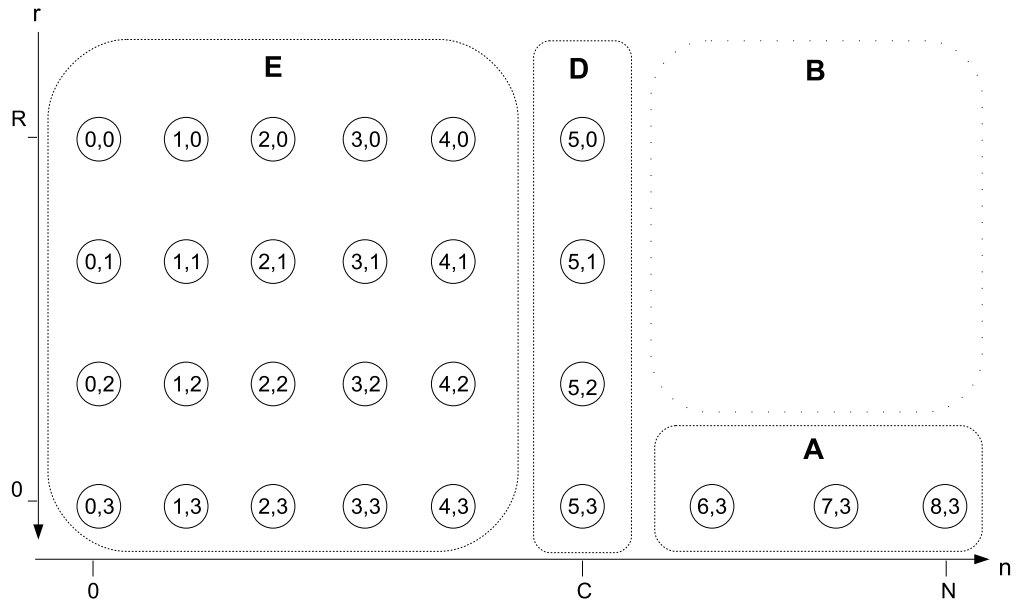


Figura 3.15: struttura e suddivisione in aree degli stati del modello con ambulanze di terze parti.

bilanciamento. La suddivisione è mostrata in Figura 3.15.

$\mathbf{A} \equiv \{\, \langle n, r \rangle \;:\; n \in 0, \ldots, N;\, r \in 0, \ldots, R;\, \mathbf{n} > \mathbf{C};\, \mathbf{r} = \mathbf{R} \,\}$
Gli appartenenti a questo insieme rispondono alle stesse considerazioni fatte per il modello visto nel Paragrafo3.2. In questi stati non si trovano in zona critica e di conseguenza non è consentito il noleggio di mezzi aggiuntivi: l'arrivo di una richiesta provoca una transizione verso sinistra lungo l'asse n mentre un completamento ne provoca una verso destra.

$\mathbf{B} \equiv \{\, \langle n, r \rangle \;:\; n \in 0, \ldots, N;\, r \in 0, \ldots, R;\, \mathbf{n} > \mathbf{C};\, \mathbf{r} < \mathbf{R} \,\}$
In questo insieme sono raccolti gli stati inammissibili dove, nonostante non ci si trovi in zona critica ($n > C$), è previsto l'utilizzo di mezzi di terzi ($r < R$). Per evitare la caduta in questi stati viene fissata a zero il loro valore di probabilità:

$$x_{\langle n,r \rangle} = 0 \qquad \forall n > C, \forall r > 0$$

con $n \in 0, \ldots, N$ ed $r \in 0, \ldots, R$.

$\mathbf{D} \equiv \{\, \langle n, r \rangle \;:\; n \in 0, \ldots, N;\, r \in 0, \ldots, R;\, \mathbf{n} = \mathbf{C} \,\}$

Gli stati appartenenti a questo insieme hanno una caratteristica comune: nonostante si trovino in zona critica con $n = C$, si ha che $n + r \geq C$ in qualsiasi caso. Questo fa sì che valgano le considerazioni fatte per arrivare all'equazione (3.17). Ciò significa che, nonostante sia teoricamente possibile servire una richiesta con un noleggio, il peso per tasso di transizione di nascita verso l'alto è sempre zero (dato da $(C - n)/C$ con $n = C$). Alla luce di questa considerazione, le ambulanze "a gettone" non vengono mai realmente impiegate, in questi stati è possibile solo la loro smobilitazione. Si veda a tal riguardo lo schema delle transizioni per l'insieme D riportato in Figura 3.16: le uniche transizioni che avvengono lungo l'asse r sono quelle di completamento del servizio con conseguente smobilitazione dei mezzi precedentemente impiegati. Le equazioni di bilanciamento per gli stati di questo insieme sono:

$$
\begin{aligned}
& x_{\langle n+1,r \rangle} \lambda + \\
& x_{\langle n,r-1 \rangle} \left(N - n + R - r + 1\right) \mu + \\
& x_{\langle n-1,r \rangle} \left(N - n + 1\right) \mu - \\
& x_{\langle n,r \rangle} \left(\left(N - n + R - r\right) \mu + \lambda\right) = 0 \qquad \text{con } n = C, r = R
\end{aligned}
$$

$$
\begin{aligned}
& x_{\langle n,r-1 \rangle} \left(N - n + R - r + 1\right) \mu + \\
& x_{\langle n-1,r \rangle} \left(N - n + 1\right) \mu - \\
& x_{\langle n,r \rangle} \left(\left(N - n + R - r\right) \mu + \lambda\right) = 0 \qquad \text{con } n = C, 0 < r < R
\end{aligned}
$$

$$
\begin{aligned}
& x_{\langle n-1,r \rangle} \left(N - n + 1\right) \mu - \\
& x_{\langle n,r \rangle} \left(\left(N - n + R - r\right) \mu + \lambda\right) = 0 \qquad \text{con } n = C, r = 0
\end{aligned}
$$

dove per tutte vale $n = C$.

$\mathbf{E} \equiv \{\, \langle n, r \rangle \;:\; n \in 0, \ldots, N;\, r \in 0, \ldots, R;\, \mathbf{n} < \mathbf{C} \,\}$

Gli stati appartenenti a questo insieme sono tutti parte della zona critica del servizio dove coesistono entrambe le possibilità di transizione elencate nell'equazioni (3.18) e (3.19). In questi stati sono previste tutte le possibili transizioni dato che è sempre consentito il noleggio per il servizio delle richieste. La struttura e le transizioni ammesse per un generico stato in zona critica sono presentate in Figura 3.17. Come si nota, le transizioni sono tutte ammissibili, sia sull'asse n delle ambulanze di proprietà (nascita verso sinistra, morte verso destra), sia

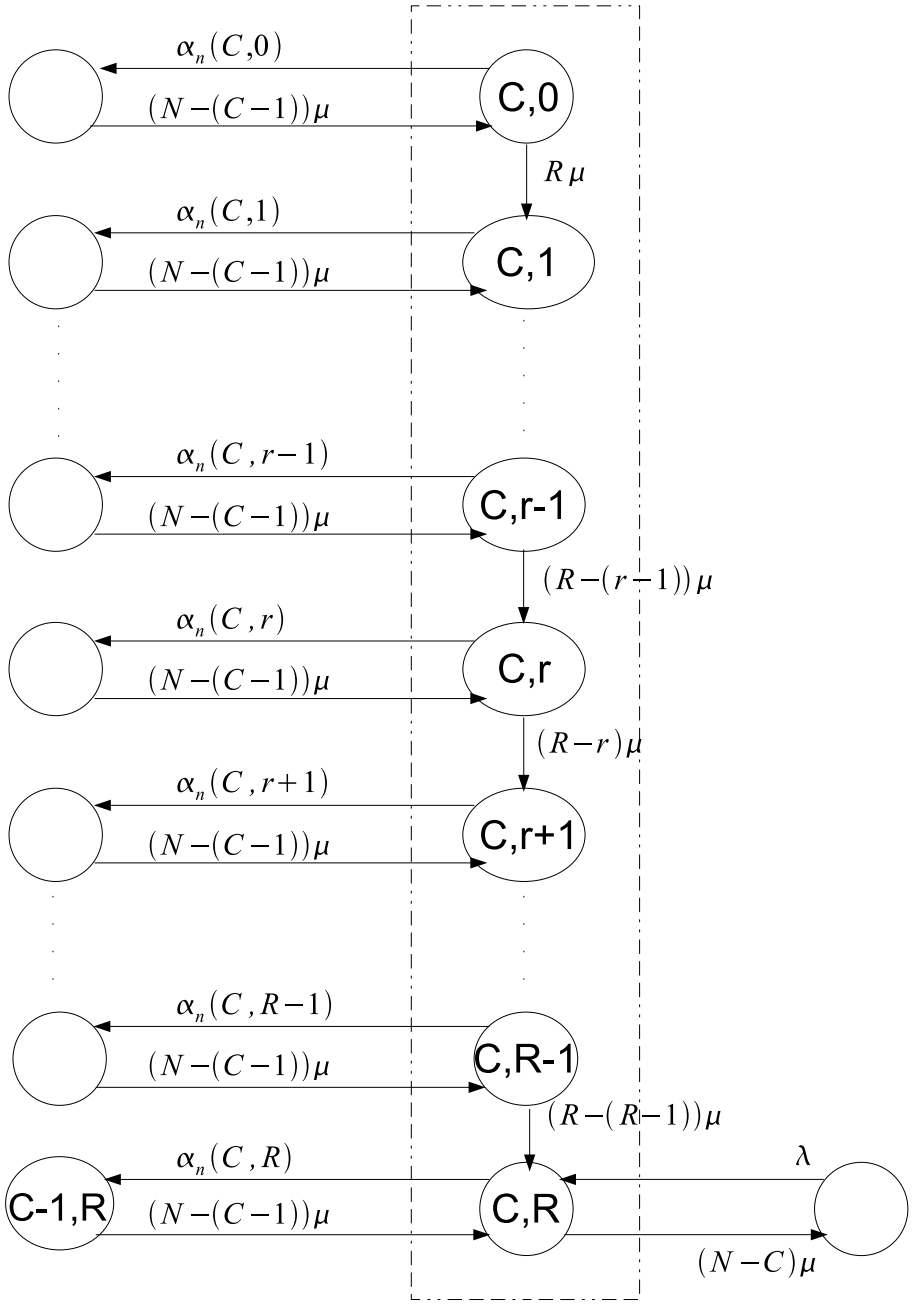


Figura 3.16: struttura e transizioni per gli stati appartenenti all'insieme D.

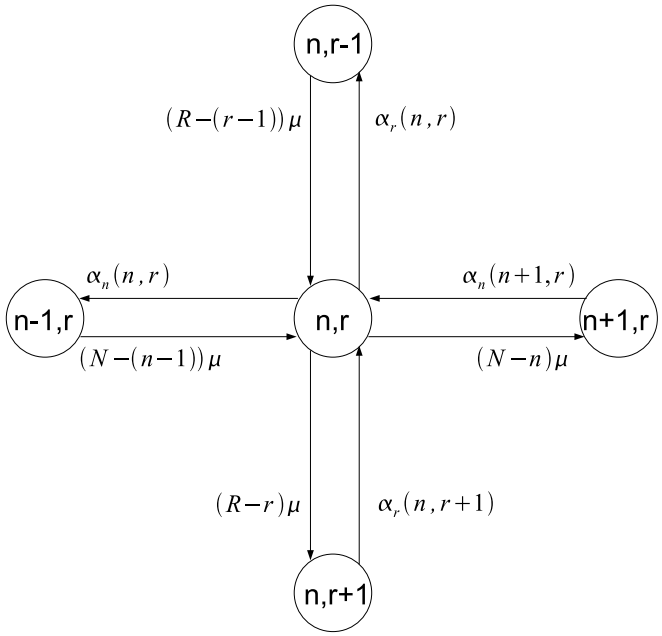


Figura 3.17: struttura e tassi delle transizioni di un generico stato in zona critica (insieme E).

lungo l'asse r dei mezzi a noleggio (nascita verso l'alto, morte verso il basso). L'equazione di bilanciamento dello stato raffigurato in Figura 3.17 è:

$$
\begin{aligned}
&x_{\langle n+1,r\rangle}\alpha_n(n+1,r)+\\
&x_{\langle n,r-1\rangle}\,(R-r+1)\,\mu+\\
&x_{\langle n-1,r\rangle}\,(N-n+1)\,\mu+\\
&x_{\langle n,r+1\rangle}\alpha_r(n,r+1)-\\
&x_{\langle n,r\rangle}\,((N-n)\,\mu+(R-r)\,\mu+\alpha_n(n,r)+\alpha_r(n,r))=0 \qquad \text{con } n<C, 0<r<R
\end{aligned}
$$

mentre, per gli altri stati dell'insieme E, i bilanciamenti risultano:

$$
\begin{aligned}
&x_{\langle n+1,r\rangle}\alpha_n(n+1,r)+\\
&x_{\langle n,r-1\rangle}\,(R-r+1)\,\mu+\\
&x_{\langle n-1,r\rangle}\,(N-n+1)\,\mu-\\
&x_{\langle n,r\rangle}\,((N-n)\,\mu+\alpha_n(n,r)+\alpha_r(n,r))=0 \qquad \text{con } n<C, r=R
\end{aligned}
$$

$$
\begin{aligned}
&x_{\langle n+1,r\rangle}\lambda+\\
&x_{\langle n-1,r\rangle}\left(N-n+1\right)\mu+\\
&x_{\langle n,r+1\rangle}\alpha_r(n,r)-\\
&x_{\langle n,r\rangle}\left(\left(N-n\right)\mu+\left(R-r\right)\mu+\lambda\right)=0 \qquad \text{con } n<C, r=0
\end{aligned}
$$

$$
\begin{aligned}
&x_{\langle n+1,r\rangle}\alpha_n(n+1,r)+\\
&x_{\langle n,r-1\rangle}\left(R-r+1\right)\mu-\\
&x_{\langle n,r\rangle}\left(\left(N-n\right)\mu+\lambda\right)=0 \qquad \text{con } n=0, r=R
\end{aligned}
$$

$$
\begin{aligned}
&x_{\langle n+1,r\rangle}\alpha_n(n+1,r)+\\
&x_{\langle n,r-1\rangle}\left(R-r+1\right)\mu+\\
&x_{\langle n,r+1\rangle}\lambda-\\
&x_{\langle n,r\rangle}\left(\left(N-n\right)\mu+\left(R-r\right)\mu+\lambda\right)=0 \qquad \text{con } n=0, 0<r<R
\end{aligned}
$$

$$
\begin{aligned}
&x_{\langle n+1,r\rangle}\alpha_n(n+1,r)+\\
&x_{\langle n,r+1\rangle}\lambda-\\
&x_{\langle n,r\rangle}\left(\left(N-n\right)\mu+\left(R-r\right)\mu\right)=0 \qquad \text{con } n=0, r=0
\end{aligned}
$$

dove per tutte vale $n<C$. Anche in questo caso è necessario aggiungere un vincolo al valore delle $x_{\langle n,r\rangle}$ probabilità per limitarne il valore complessivo a uno.

3.4.3 Risultati

L'approccio alla risoluzione del sistema risultante è lo stesso utilizzato per i modelli visti in precedenza: formalizzazione in *GNU MathProg* e risoluzione tramite *GLPK*. I dati risultanti dalla soluzione permettono di analizzare due aspetti distinti dell'andamento delle prestazioni: il primo è il livello di servizio così come visto per gli altri modelli. Esso permette di valutare le frazioni del tempo in cui il sistema si trova in un dato livello di occupazione dei mezzi. Il secondo aspetto riguarda il noleggio dei mezzi di terze parti: considerando i valori delle probabilità $x_{\langle n,r\rangle}$ "per righe" del modello visto come una matrice (si veda in Figura 3.15), si possono ottenere le frazioni di tempo sul totale della fascia nelle quali si ha un dato livello di utilizzo delle ambulanze "a gettone". Questa distribuzione è ricavata sommando le

probabilità su ogni r, numero di mezzi di terzi disponibili:

$$x_{\langle r \rangle} = \sum_{n=0}^{N} x_{\langle n,r \rangle} \qquad \forall r \in 0, \ldots, R$$

mentre il livello di servizio sommando per colonne

$$x_{\langle n \rangle} = \sum_{r=0}^{R} x_{\langle n,r \rangle} \qquad \forall n \in 0, \ldots, N$$

ottenendo così gli $x_{\langle n \rangle}$ valori. In Figura 3.18 sono riportati i risultati del livello di servizio (grafico di sinistra) e dello stato dell'utilizzo dei mezzi a noleggio (grafico di destra); i dati sono stati calcolati risolvendo il sistema lineare relativo a dimensioni crescenti della flotta di proprietà. Come è prevedibile, ad un numero maggiore di ambulanze proprie, corrisponde un minore utilizzo di quelle "a gettone". Procedendo in questo modo, il decisore potrebbe dimensionare la flotta in funzione di un utilizzo accettabile dei mezzi di terzi. In modo analogo, il decisore potrebbe voler operare come mostrato in Figura 3.19 ovvero, mantenendo costante il valore dei parametri N (ambulanze di proprietà) ed R (ambulanze di "a gettone"), valutare come il *mix* di parametri sotto esame si ripercuote sulle prestazioni del servizio lungo fasce orarie di criticità crescente (λ in aumento).

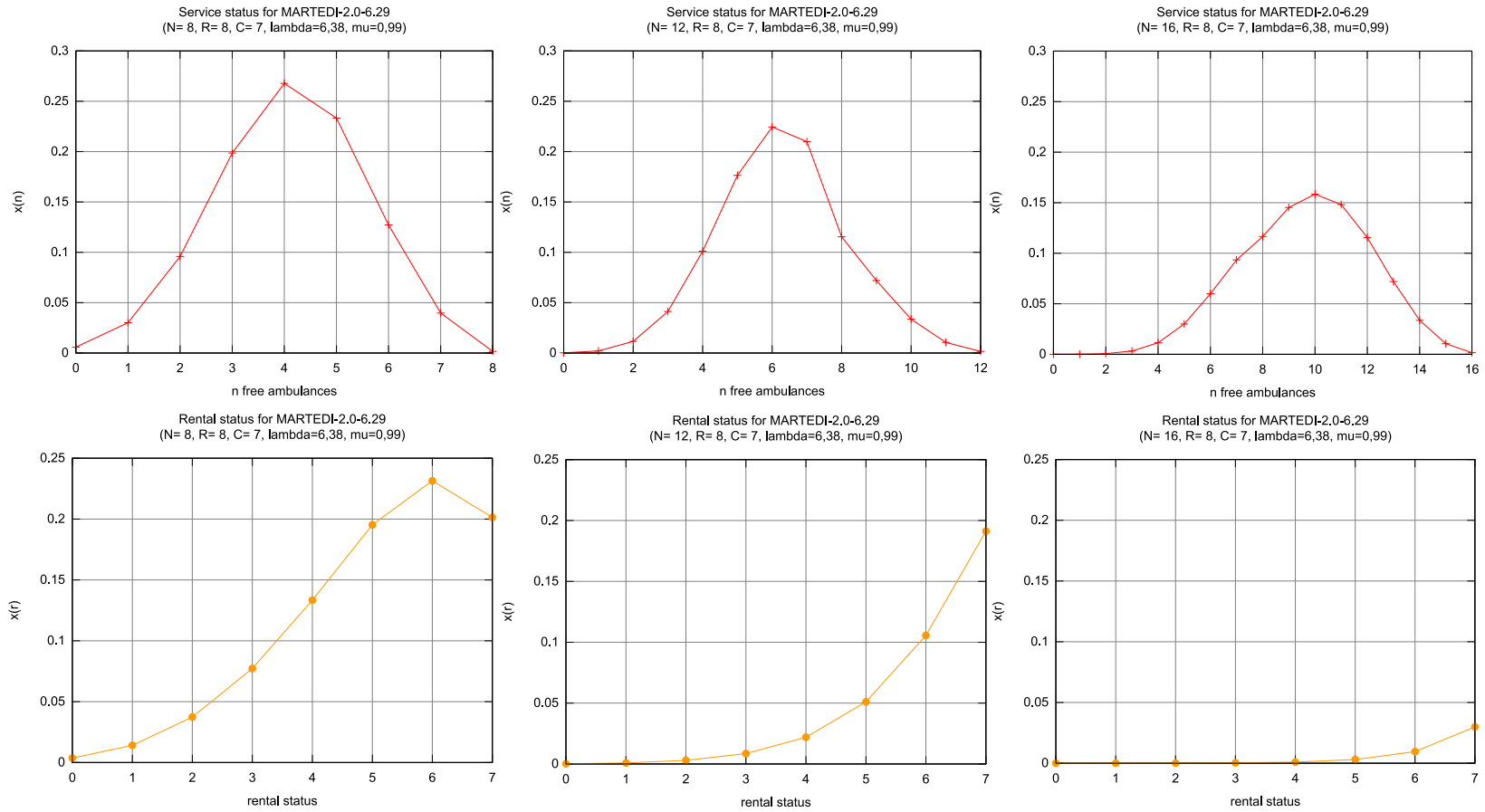


Figura 3.18: grafici del livello di servizio e stato dell'utilizzo di mezzi di terzi al variare della dimensione della flotta di proprietà.

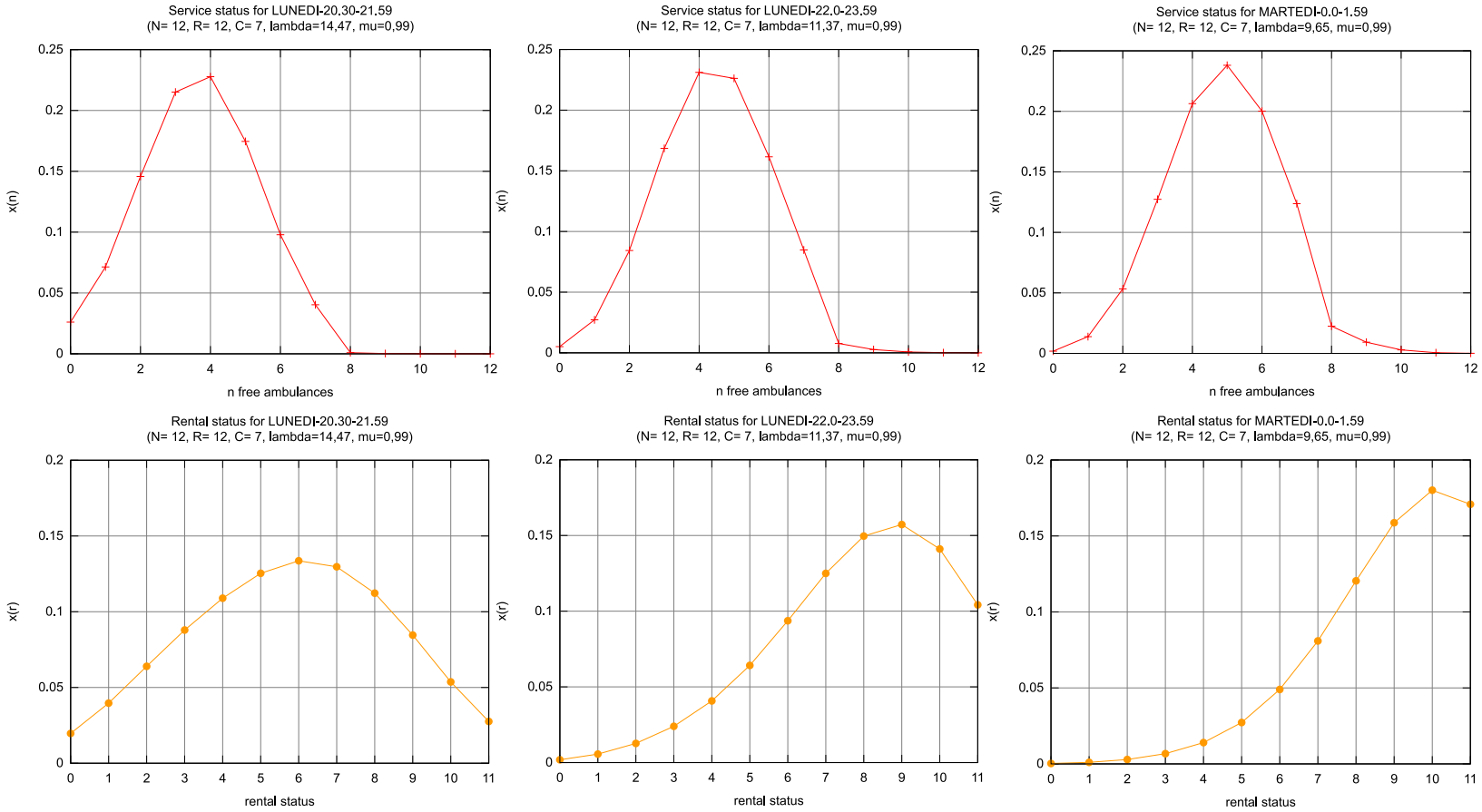


Figura 3.19: grafici del livello di servizio e stato dell'utilizzo di mezzi di terzi al variare della fascia oraria considerata (variazione di λ).

3.5 Un modello per la valutazione complessiva delle variabili decisionali

Il quarto approccio al problema decisionale ha previsto la realizzazione di un modello che permettesse di considerare in un unico sistema tutte le variabili viste nei casi precedenti. Questa terza estensione è pensata per fornire al decisore la facoltà di valutare le interazioni fra i diversi aspetti delle performance del Servizio "118", analizzare quali effetti un dato mix di variabili ha sull'andamento del livello di servizio ma, nel contempo, anche le ripercussioni sull'attesa da parte delle richieste non urgenti o sulla richiesta di utilizzo di mezzi a noleggio. Combinare i modelli visti nei Paragrafo 3.3 e 3.4 ha portato ad una estensione completa con la quale è possibile valutare, utilizzando un unico strumento, l'effetto che tutte le variabili decisionali hanno sul funzionamento del servizio.

Questo modello considera quindi sia le variabili previste dall'accodamento delle richieste "verdi" (Paragrafo 3.3), sia quelle previste dalla gestione delle ambulanze "a gettone" (Paragrafo 3.4):

- N numero di ambulanze di proprietà dell'azienda ospedaliera;
- R numero totale di mezzi di terze parti disponibili;
- C numero di zone del territorio da servire;
- K livello di guardia per il servizio delle chiamate non urgenti.

Dato che in questo caso sono previste entrambe le problematiche, si è deciso di considerare come prioritario l'utilizzo di mezzi a noleggio in quanto questo produce un costo aggiuntivo ma non un degrado del servizio. L'ultima risorsa da prendere in considerazione è la messa in attesa delle chiamate "verdi": se nonostante l'impiego di mezzi aggiuntivi non si riesce a tenere il sistema fuori dalla zona critica (il numero di mezzi disponibili è minore di K), allora è consentito procedere all'accodamento.

3.5.1 Il modello

Ogni stato di questo modello è caratterizzato da una tripla di indici $\langle n, r, k\rangle$: $n \in 0, \ldots, N$ numero di ambulanze proprie rimaste libere, $r \in 0, \ldots, R$ numero di mezzi a noleggio disponibili all'utilizzo, $k \in 0, \ldots, K$ numero di chiamate non urgenti in attesa in coda. In ognuno

di questi stati, il numero complessivo di mezzi disponibili è dato da $n + r$. In Figura 3.20 è

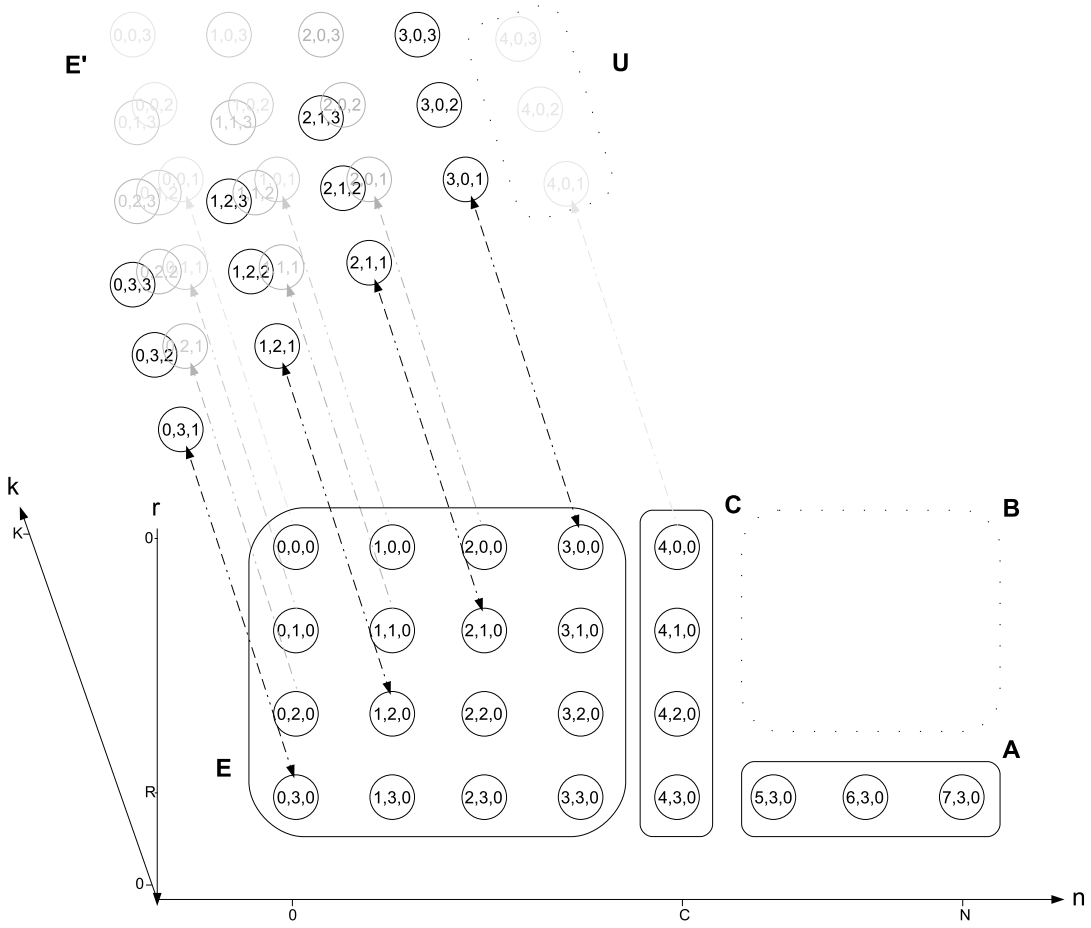


Figura 3.20: rappresentazione schematica della struttura del modello combinato.

rappresentata la struttura tridimensionale degli stati ottenuta con $N = 7$, $R = 3$, $C = 4$ e $K = 3$. Come si nota dallo schema, il modello costruito sulle prime due dimensioni, l'asse n e l'asse r, ricorda quanto visto per il modello di Figura 3.15 (a p. 55). Il funzionamento delle transizioni su queste due dimensioni è di fatto lo stesso: dato C numero delle zone di competenza che devono essere servite (coperte da almeno un mezzo), la zona critica per il noleggio si ha solo quando $n \leq C$. Come si nota, l'utilizzo delle ambulanze "a gettone" (che prevede transizioni lungo l'asse r) è consentito solo per gli stati che sono entrati nella prima zona critica. La differenza sostanziale introdotta da questa estensione è la coesistenza delle due zone critiche: sia quella relativa al noleggio, sia quella relativa alle chiamate non urgenti.

La seconda di queste, sicuramente più problematica, ha inizio quando l'insieme di ambulanze libere scende al di sotto di una dimensione di guardia data (K); dato che il numero di mezzi liberi è $n + r$, la zona di guardia si ha negli stati dove $n + r \leq K$. Come si può notare anche dalla Figura 3.20, gli stati relativi alla coda per le richieste non urgenti sono disposti lungo l'asse k, la terza dimensione lungo la quale avvengono le transizioni di svuotamento e riempimento della coda d'attesa. Gli stati "di frontiera" in cui si ha che $n + r = K$ formano una superficie lungo la quale è possibile scendere (svuotare la coda per tornare al "piano" $k = 0$ dei soli assi n ed r, senza codici verdi in attesa) o salire mettendo codici verdi in attesa. Ad ogni modo, come succedeva per il modello con attesa per i codici verdi (Paragrafo 3.3), è possibile svuotare la coda, servendo richieste non urgenti in attesa, solo nell'ambito di questi stati di frontiera dove il numero di mezzi disponibili eguaglia la soglia di guardia K; in tutti gli altri stati dove $n+r < K$ e $k > 0$ non sono previste transizioni verso il basso lungo l'asse k, non è quindi possibile ritornare a valori inferiori di k fintanto che il numero di ambulanze libere rimane al di sotto della soglia critica.

Anche per l'analisi di questo modello è stata adottata una tecnica di "partizionamento" in insiemi come visto per le varianti precedenti. Gli insiemi individuati, mostrati in Figura 3.20, si distinguono per differenti caratteristiche delle equazioni di bilanciamento.

$\mathbf{A} \equiv \{\, \langle n, r, k \rangle \;:\; n \in 0, \ldots, N;\, r \in 0, \ldots, R;\, k \in 0, \ldots, K;\, \mathbf{n} > \mathbf{C};\, \mathbf{r} = \mathbf{R};\, \mathbf{k} = \mathbf{0} \,\}$
Gli stati appartenenti a questo insieme non si trovano in nessuna zona critica: le ambulanze di proprietà sono sufficienti a coprire qualsiasi zona e non è consentito l'accodamento dei codici verdi. Questi stati rispondono alle stesse considerazioni fatte per l'insieme omonimo del modello visto nel Paragrafo 3.4: le uniche transizioni ammesse sono quelle lungo l'asse n con l'arrivo di una richiesta verso sinistra ($n - 1$), il completamento verso destra ($n + 1$). I tassi di transizione usano valori indifferenziati λ^{tot} e μ^{tot} per i processi di nascita e morte.

$\mathbf{B} \equiv \{\, \langle n, r, k \rangle \;:\; n \in 0, \ldots, N;\, r \in 0, \ldots, R;\, k \in 0, \ldots, K;\, \mathbf{n} > \mathbf{C};\, \mathbf{r} < \mathbf{R};\, \}$
Questa partizione rappresenta lo stesso insieme di inammissibilità visto per il modello di Paragrafo 3.4: in questi stati, nonostante non ci si trovi in zona critica, si ha l'utilizzo di mezzi di terze parti ($r < R$). L'inammissibilità di questi stati riguarda quindi l'asse r.

$\mathbf{U} \equiv \{\, \langle n, r, k \rangle \;:\; n \in 0, \ldots, N;\, r \in 0, \ldots, R;\, k \in 0, \ldots, K;\, \mathbf{n} + \mathbf{r} > \mathbf{K};\, \mathbf{k} > \mathbf{0} \,\}$
Gli stati che si trovano in questa partizione rappresentano il principale insieme di inammis-

sibilità: nonostante non si trovino in alcuna situazione critica ($n + r > K$), essi prevedono l'accodamento di richieste non urgenti ($k > 0$). Ciò significa che, nonostante rimangano validi per le dimensioni n ed r, questi stati sono non ammessi per la dimensione k, la lunghezza della coda. Ovviamente questo comportamento non è contemplato: come avviene per ogni stato non ammissibile, l'utilizzo è inibito ponendo il suo valore di probabilità a zero. In questo modo viene impedito l'utilizzo della terza dimensione (asse k, accodamento richieste).

$\mathbf{D} \equiv \{\, \langle n, r, k\rangle \;:\; n \in 0, \ldots, N;\, r \in 0, \ldots, R;\, k \in 0, \ldots, K;\, \mathbf{n} = \mathbf{C};\, \mathbf{k} = \mathbf{0};\, \mathbf{n} + \mathbf{r} > \mathbf{K} \,\}$
Gli stati appartenenti a questa partizione si trovano in zona critica per quanto riguarda il noleggio di mezzi esterni. Si tenga presente però che le considerazioni fatte per i tassi di transizione in zona critica (Paragrafo 3.4) continuano a valere, così come le osservazioni introdotte per arrivare all'equazione (3.17) (a pagina 53). Ciò significa che, nonostante sia teoricamente possibile servire una richiesta con un noleggio, il peso per tasso di transizione di nascita lungo l'asse r è sempre nullo: le ambulanze "a gettone" non vengono mai realmente impiegate ed in questi stati è possibile solo la loro smobilitazione. Le equazioni di bilanciamento per questa partizione, differenziate per gli stati di bordo, sono:

$$
\begin{aligned}
& x_{\langle n+1,r,k\rangle}\lambda^{tot} + \\
& x_{\langle n,r-1,k\rangle}\,(N-n+R-r+1)\,\mu^{tot} + \\
& x_{\langle n-1,r,k\rangle}\,(N-n+1)\,\mu^{tot} - \\
& x_{\langle n,r,k\rangle}\left((N-n+R-r)\,\mu^{tot} + \lambda^{tot}\right) = 0 \qquad \text{con } r = R
\end{aligned}
$$

$$
\begin{aligned}
& x_{\langle n,r-1,k\rangle}\,(N-n+R-r+1)\,\mu^{tot} + \\
& x_{\langle n-1,r,k\rangle}\,(N-n+1)\,\mu^{tot} - \\
& x_{n,r,k}\left((N-n+R-r)\,\mu^{tot} + \lambda^{tot}\right) = 0 \qquad \text{con } 0 < r < R
\end{aligned}
$$

$$
\begin{aligned}
& x_{\langle n-1,r,k\rangle}\,(N-n+1)\,\mu^{tot} - \\
& x_{\langle n,r,k\rangle}\left((N-n+R-r)\,\mu^{tot} + \lambda^{tot}\right) = 0 \qquad \text{con } r = 0
\end{aligned}
$$

Inoltre non è previsto l'accodamento dei codici verdi: il valore di k fissato a zero pone di fatto questo insieme equivalente all'omonimo già presentato nell'ambito del Paragrafo 3.4.

$\mathbf{D'} \equiv \{\, \langle n, r, k\rangle \;:\; n \in 0, \ldots, N;\; r \in 0, \ldots, R;\; k \in 0, \ldots, K;\; \mathbf{n} = \mathbf{C};\; \mathbf{n} + \mathbf{r} \leq \mathbf{K} \,\}$

Gli stati di questa partizione si trovano in entrambe le zone critiche: sono quindi consentiti sia il noleggio (anche se, come spiegato per l'insieme D, i pesi per le transizioni di noleggio su r sono nulli) che l'accodamento dei codici verdi in arrivo (in quanto $n + r \leq K$). Un esempio schematico di stato è riportato in Figura 3.21. Come si vede dalla rappresentazione, è necessario differenziare la struttura delle transizioni a seconda del valore di k. Nel primo caso presentato (3.21(a)) si ha che $n+r = K$: ciò significa che lo stato $\langle n, r, 0\rangle$ è sul piano di confine con la zona critica per le richieste "verdi". L'equazione di bilanciamento per questa prima situazione è:

$$
\begin{aligned}
& x_{\langle n,r-1,k\rangle}\,(N - n + R - r + 1)\,\mu^{tot} + \\
& x_{\langle n-1,r,k\rangle}\,(N - n + 1)\,\mu^{tot} + \\
& x_{\langle n,r,k+1\rangle}\,(N - n + R - r)\,\mu^{tot} - \\
& x_{\langle n,r,k\rangle}\left((N - n + R - r)\,\mu^{tot} + \lambda^{u} + \lambda^{v}\right) = 0 \qquad \text{con } 0 < r < R,\; k = 0
\end{aligned}
$$

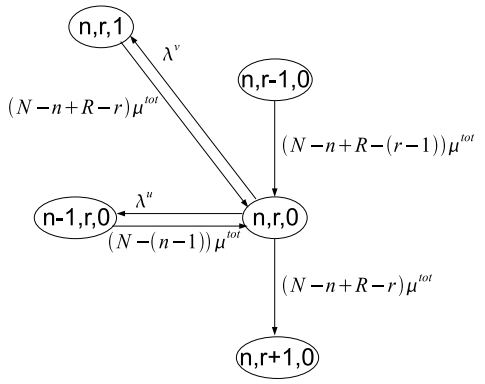
Come si può notare, in questa situazione si ha anche che $k = 0$: gli archi di transizione lungo l'asse k sono bidirezionali ma solo per $k + 1$, non per $k - 1$ (in quanto $k = 0$).

Nel secondo caso (3.21(b)) continua ad essere $n+r = K$ ma $k > 0$. Questo significa che lo stato $\langle n, r, k\rangle$ è sempre sul confine della zona critica ma non è più al livello di partenza dove $k = 0$. L'equazione di bilanciamento per questa seconda situazione è:

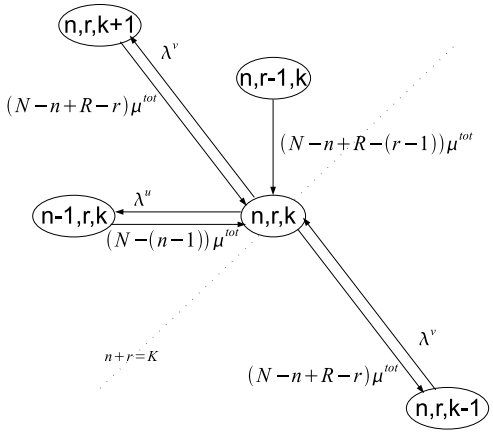
$$
\begin{aligned}
& x_{\langle n,r-1,k\rangle}\,(N - n + R - r + 1)\,\mu^{tot} + \\
& x_{\langle n-1,r,k\rangle}\,(N - n + 1)\,\mu^{tot} + \\
& x_{\langle n,r,k+1\rangle}\,(N - n + R - r)\,\mu^{tot} + \\
& x_{\langle n,r,k-1\rangle}\lambda^{v} - \\
& x_{\langle n,r,k\rangle}\left((N - n + R - r)\,\mu^{tot} + \lambda^{u} + \lambda^{v}\right) = 0 \qquad \text{con } 0 < r < R,\; k > 0
\end{aligned}
$$

Le transizioni lungo l'asse k sono bidirezionali (è possibile sia accodare che servire) e sono presenti sia per/da $k + 1$ che $k - 1$. É interessante notare che, a differenza del caso con $k = 0$, qui la transizione verso $\langle n, r + 1, k\rangle$ è assente: essendo lo stato $\langle n, r, k\rangle$ posizionato sul piano verticale di confine, gli stati del tipo $\langle n, r + 1, k\rangle$ appartengono alla partizione di inammissibilità U.

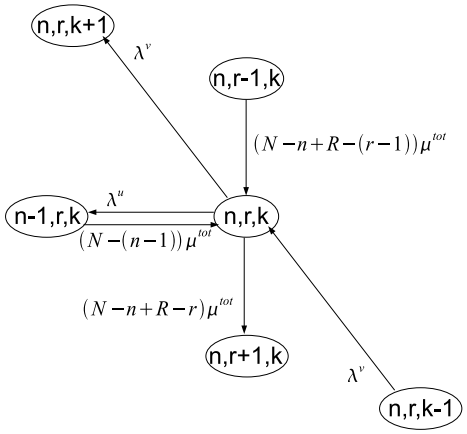
Il terzo caso presentato (3.21(c)) riguarda gli stati dove $n + r < K$ e $k > 0$. In essi



(a) $\mathbf{n = C}$, $\mathbf{n + r = K}$, $\mathbf{k = 0}$



(b) $\mathbf{n = C}$, $\mathbf{n + r = K}$, $\mathbf{k > 0}$



(c) $\mathbf{n = C}$, $\mathbf{n + r < K}$, $\mathbf{k > 0}$

Figura 3.21: struttura e transizioni per un generico stato appartenente all'insieme D' ($n = C$; $n + r \leq K$).

sono accodate k richieste non urgenti e, come visto nel modello di Paragrafo 3.3, l'unica transizione ammessa lungo k è l'accodamento.

$$
\begin{aligned}
& x_{\langle n,r-1,k\rangle}\,(N-n+R-r+1)\,\mu^{tot}+ \\
& x_{\langle n-1,r,k\rangle}\,(N-n+1)\,\mu^{tot}+ \\
& x_{\langle n,r,k-1\rangle}\lambda^{v}- \\
& x_{\langle n,r,k\rangle}\left((N-n+R-r)\,\mu^{tot}+\lambda^{u}+\lambda^{v}\right)=0 \qquad \text{con } 0<r<R,\ k>0
\end{aligned}
$$

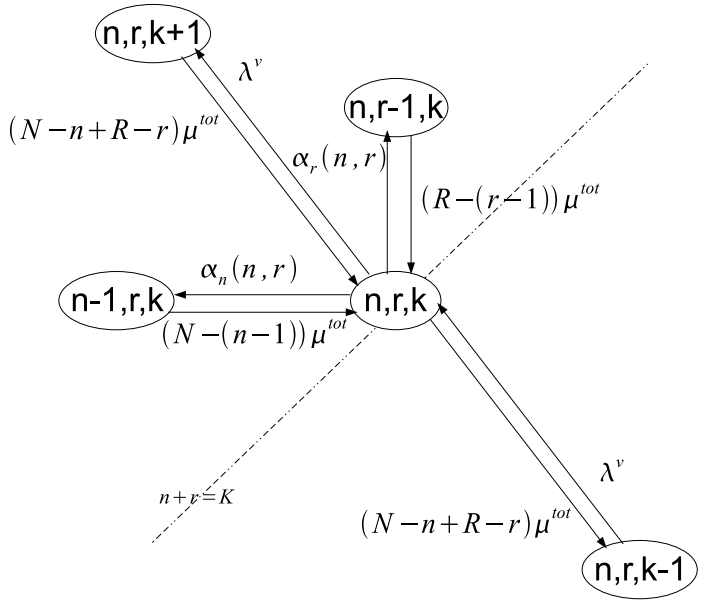
In quest'ultima situazione, le uniche transizioni di servizio ammesse hanno tasso λ^u (solo urgenze, codici gialli e rossi). Le rimanenti, aventi tasso λ^v, portano ad un accodamento del cliente.

$\mathbf{E} \equiv \{\, \langle n,r,k\rangle \;:\; n \in 0,\ldots,N;\, r \in 0,\ldots,R;\, k \in 0,\ldots,K;\, \mathbf{n} < \mathbf{C};\, \mathbf{k} = \mathbf{0};\, \mathbf{n}+\mathbf{r} > \mathbf{K} \,\}$
Gli stati che ricadono in questa partizione sono considerati in zona critica solo per quanto riguarda il noleggio di mezzi aggiuntivi: dato che $n < C$, è previsto il ricorso alle ambulanze "a gettone" secondo le considerazioni fatte per l'insieme E del modello di Paragrafo 3.4. Inoltre, essendo fuori dalla zona critica per quanto riguarda le chiamate non urgenti ($n+r > K$), il valore di k è obbligatoriamente fissato a zero. La generica equazione di bilanciamento per questa partizione risulta:

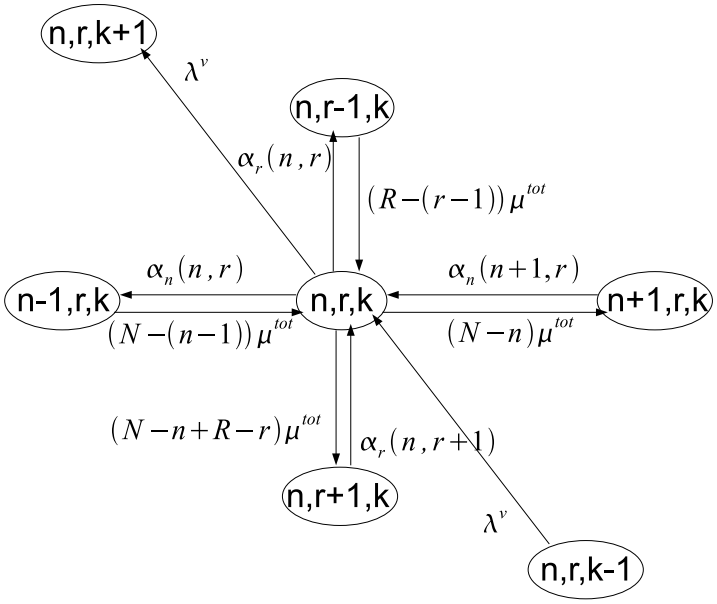
$$
\begin{aligned}
& x_{\langle n+1,r,k\rangle}\alpha_n(n+1,r)+ \\
& x_{\langle n,r-1,k\rangle}\,(R-r+1)\,\mu^{tot}+ \\
& x_{\langle n-1,r,k\rangle}\,(N-n+1)\,\mu^{tot}+ \\
& x_{\langle n,r+1,k\rangle}\alpha_r(n,r+1)- \\
& x_{\langle n,r,k\rangle}\left((N-n)\,\mu^{tot}+(R-r)\,\mu^{tot}+\alpha_n(n,r)+\alpha_r(n,r)\right)=0
\end{aligned}
$$

Alla luce di queste considerazioni si nota come le caratteristiche degli stati appartenenti a questo insieme sono analoghe a quelle degli stati dell'insieme omonimo visto nell'ambito del modello per la valutazione dell'utilizzo di mezzi "a gettone" (Paragrafo 3.4).

$\mathbf{E'} \equiv \{\, \langle n,r,k\rangle \;:\; n \in 0,\ldots,N;\, r \in 0,\ldots,R;\, k \in 0,\ldots,K;\, \mathbf{n} < \mathbf{C};\, \mathbf{n}+\mathbf{r} \leq \mathbf{K} \,\}$
Gli stati di questo insieme sono considerati in zona critica da entrambi i criteri: con $n < C$



(a) $\mathbf{n+r=K,\ k>0}$



(b) $\mathbf{n+r<K,\ k>0}$

Figura 3.22: struttura e transizioni per un generico stato appartenente all'insieme E' ($n < C;\ n+r \geq K$).

è consentito l'utilizzo di mezzi di terze parti per il servizio delle zone di competenza, con $n + r \leq K$ è permesso l'accodamento delle richieste identificate come codici verdi.

In Figura 3.22 è riportata una rappresentazione schematica delle transizioni previste per questi stati. Come si può notare, anche in questo caso si è resa necessaria una distinzione dello schema a seconda dei valori di n ed r. Nella prima situazione presentata (3.22(a)) si ha che $n+r = K$: lo stato $\langle n, r, k\rangle$ si trova quindi sul piano verticale di confine con la zona critica per le chiamate non urgenti. Si noti che le transizioni lungo l'asse k sono bidirezionali, è consentito sia l'accodamento che la messa in attesa dei codici verdi. Nello schema riportato si ha anche che $k > 0$ con conseguente presenza di richieste non urgenti in attesa e transizione verso $\langle n, r, k-1\rangle$ consentita (dato che la coda non è vuota, è possibile servire un cliente in attesa). Come visto per il modello relativo al ricorso ai mezzi "a gettone" (Paragrafo 3.4), anche in questa versione i tassi delle transizioni relative al noleggio di mezzi aggiuntivi sono funzione del valore degli indici n ed r stessi. Continuano a valere le considerazioni fatte in precedenza che hanno portato a definire il tasso di transizione $\alpha_n(n, r)$ nell'equazione (3.18) ed $\alpha_n(n, r)$ nella (3.19) (a pagina. 53). L'equazione di bilanciamento per questo caso risulta

$$
\begin{aligned}
& x_{\langle n,r,k-1\rangle}\lambda^v + \\
& x_{\langle n,r-1,k\rangle}\left(R - r + 1\right)\mu^{tot} + \\
& x_{\langle n-1,r,k\rangle}\left(N - n + 1\right)\mu^{tot} + \\
& x_{\langle n,r,k+1\rangle}\left(N - n + R - r\right)\mu^{tot} - \\
& x_{\langle n,r,k\rangle}\left(\alpha_n(n,r) + \alpha_r(n,r) + \lambda^v + \left(N - n + R - r\right)\mu^{tot}\right) = 0
\end{aligned}
$$

dove si possono ritrovare, con segno positivo, i tassi delle transizioni in ingresso e, con segno negativo, quelli per le transizioni in uscita.

Nella seconda situazione (Figura 3.22(b)) è riportato lo schema relativo ad uno stato dove $n+r = K$. Si può notare come, rispetto alla situazione precedente, le transizioni lungo l'asse k non sono bidirezionali, in questi stati non è infatti possibile servire i clienti non urgenti lasciati in attesa. Come spiegato in occasione del modello relativo all'accodamento dei codici verdi, se il sistema si trova in zona critica ma non sul confine, ogni ambulanza che torna ad essere disponibile deve essere utilizzata per il servizio di clienti urgenti o lasciata in stato di disponibilità. È per questo motivo che le transizioni relative all'arrivo di codici verdi provocano un accodamento con tasso λ^v. L'equazione di bilanciamento per gli stati di questo

tipo risulta

$$
\begin{aligned}
& x_{\langle n,r,k-1\rangle}\lambda^{v}+ \\
& x_{\langle n,r+1,k\rangle}\alpha_r(n,r+1)+ \\
& x_{\langle n+1,r,k\rangle}\alpha_n(n+1,r)+ \\
& x_{\langle n,r-1,k\rangle}\left(R-r+1\right)\mu^{tot}+ \\
& x_{\langle n-1,r,k\rangle}\left(N-n+1\right)\mu^{tot}- \\
& x_{\langle n,r,k\rangle}\left(\alpha_n(n,r)+\alpha_r(n,r)+\lambda^{v}+\left(N-n+R-r\right)\mu^{tot}+\left(N-n\right)\mu^{tot}\right)=0
\end{aligned}
$$

dove i tassi delle transizioni in ingresso hanno segno positivo, i tassi in uscita segno negativo. Per completare il sistema è stato aggiunto il vincolo di normalizzazione dei valori di tutte le probabilità:

$$\sum_{n=0}^{N}\sum_{r=0}^{R}\sum_{k=0}^{K}x_{\langle n,r,k\rangle}=1$$

3.5.2 Risultati

L'approccio alla soluzione del sistema, ovvero la sua formalizzazione in linguaggio *GNU MathProg*, è lo stesso utilizzato in precedenza. L'utilizzo di questo modello fornisce controllo su quattro variabili differenti: il decisore potrebbe voler agire su ognuna di queste con opportune variazioni e valutare la conseguente risposta del sistema. In Figura 3.23 sono riportati i grafici relativi alla risposta del sistema al variare della dimensione della flotta di ambulanze di proprietà (il parametro N). I dati di ogni test sono riportati *verticalmente*: per ogni colonna si ha il grafico del livello di servizio, il grafico dell'utilizzo dei mezzi di terze parti e per ultimo quello relativo all'utilizzo della coda d'attesa per le richieste non urgenti. Per la prova in questione è stata considerata una fascia critica con un elevato carico ($\lambda^{tot} = 18,24$). Come si può osservare, il comportamento del modello al crescere di N è sensato: con $N = 10$ si ha un intensivo utilizzo di mezzi "a gettone" e la frazione del tempo in cui si ha la coda piena è significativa. Al crescere della flotta con $N = 15$, le performance del servizio migliorano, si ha una diminuzione del ricorso a mezzi aggiuntivi ed un drastico calo della probabilità di avere dei codici verdi in coda. Aumentando ancora il valore di N, nel terzo caso portato a $N = 20$, si nota come il massimo della curva del livello di servizio si sposti a destra, così come avviene per quella relativa alle probabilità di noleggio (massimo in corrispondenza dei nove mezzi "a gettone" liberi rispetto ai tre del test precedente). An-

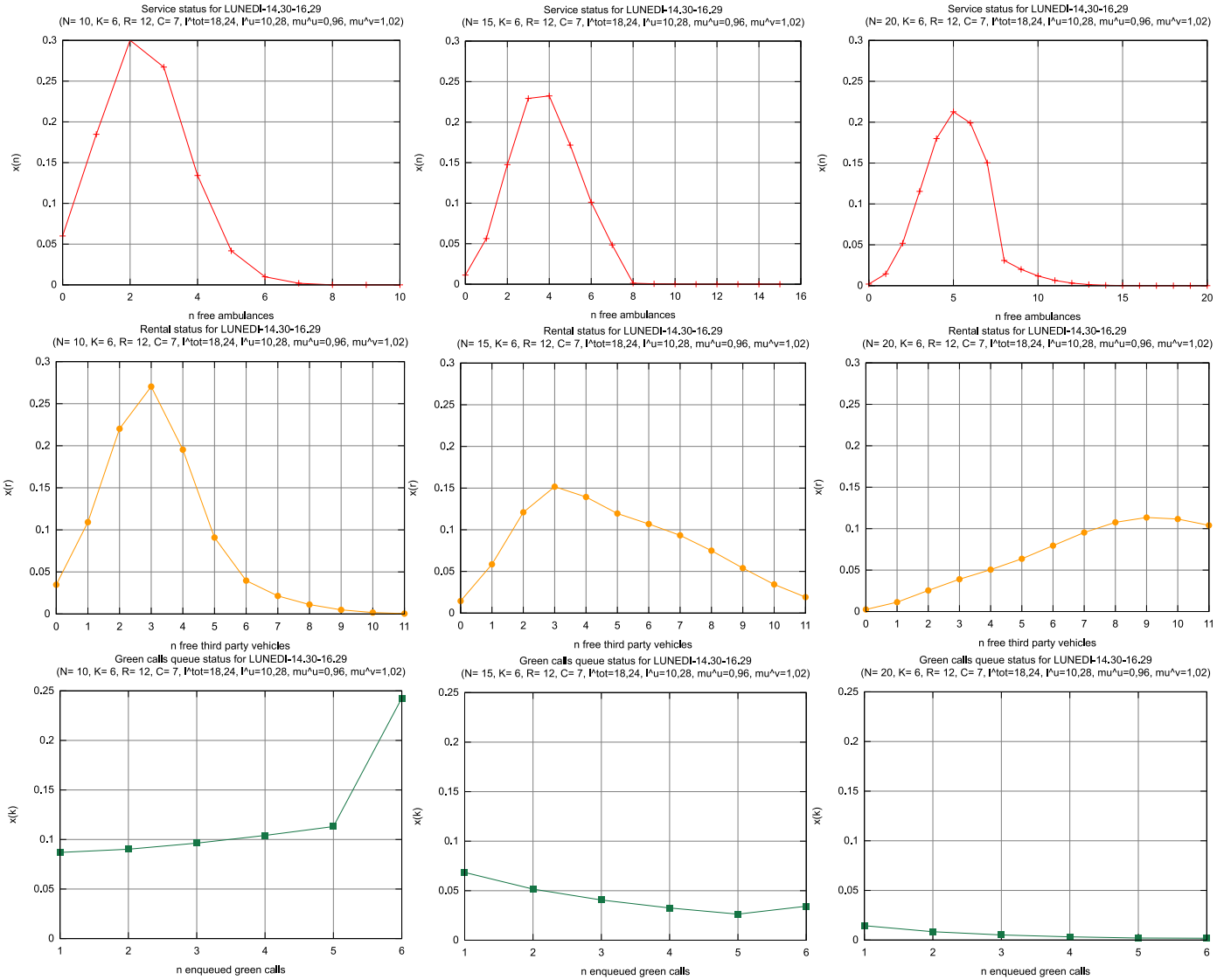


Figura 3.23: grafici del comportamento del sistema combinato al variare della dimensione della flotta di proprietà (parametro N).

che in concomitanza di quest'ultimo incremento di flotta, la necessità di accodare richieste non urgenti diminuisce raggiungendo valori ancora più bassi. Il decisore potrebbe valutare la situazione preventivata dal modello e, facendo variare il valore di N, decretarne o meno l'accettabilità. In Figura 3.24 sono riportati i risultati relativi alla performance del sistema al variare del numero di mezzi "a gettone" disponibili per il noleggio. Come si può intuire dalla priorità data all'accodamento dei codici verdi, l'aumento del parametro R non porta ad un immediato calo di utilizzo dei mezzi di proprietà ma, più verosimilmente, produce una sensibile diminuzione dell'accodamento. Si passa quindi da una prima situazione con $R = 5$, ed una significativa frazione del tempo totale della fascia con ricorso all'attesa dei clienti "verdi", ad un secondo caso con $R = 10$ dove la diminuzione delle probabilità di accodamento è significativa. Nell'ultima situazione, dove $R = 15$, si nota come il sistema privilegi lo sfruttamento dell'incrementata capacità della flotta di terzi al ricorso all'accodamento. In quest'ultima situazione la probabilità di avere clienti in attesa è ridotta a zero. Considerando questo tipo di scenario, il decisore potrebbe procedere incrementando l'utilizzo dei mezzi a noleggio per ottenere un utilizzo accettabile della coda d'attesa. Il terzo approccio all'analisi delle prestazioni del sistema è basato sulla variazione della soglia di allarme per le richieste non urgenti. I risultati ottenuti dai test sono riportati in Figura 3.25. Dai grafici si nota come un aumento della soglia di allarme, che passa da $K = 6$ a $K = 8$ e $K = 10$, porti ad un leggero aumento della probabilità totale di utilizzo della coda d'attesa e ad una leggera diminuzione del ricorso al noleggio di mezzi aggiuntivi. Questo comportamento è ragionevole: l'aumento del minimo insieme di mezzi che devono essere disponibili per non entrare in zona critica porta ad avere un numero crescente di ambulanze che non vengono utilizzate per essere lasciate disponibili per il servizio delle urgenze. In Figura 3.26 sono presentati i risultati ottenuti variando la fascia oraria di funzionamento del servizio (i parametri λ^{tot} e λ^{u}). Nell'esempio si passa da una fascia oraria di medio carico ($\lambda^{tot} = 9,24$ e $\lambda^{u} = 4,85$) ad una con un carico molto leggero ($\lambda^{tot} = 6,26$ e $\lambda^{u} = 4,85$) per arrivare infine ad una fascia altamente critica ($\lambda^{tot} = 18,24$ e $\lambda^{u} = 12,64$). Come è ragionevole aspettarsi, all'aumentare del carico al quale è sottoposto, il sistema risponde con un necessario incremento dell'utilizzo dei mezzi "a gettone" e con un inevitabile aumento del ricorso all'attesa per i codici verdi. Questo approccio può essere utilizzato dal decisore che, una volta fissati i valori delle altre variabili, vuole valutare la risposta del sistema allo scorrere del tempo. Egli potrebbe analizzare il comportamento dato dal *mix* di valori scelti a fronte di situazioni più o meno critiche e valutarne l'accettabilità.

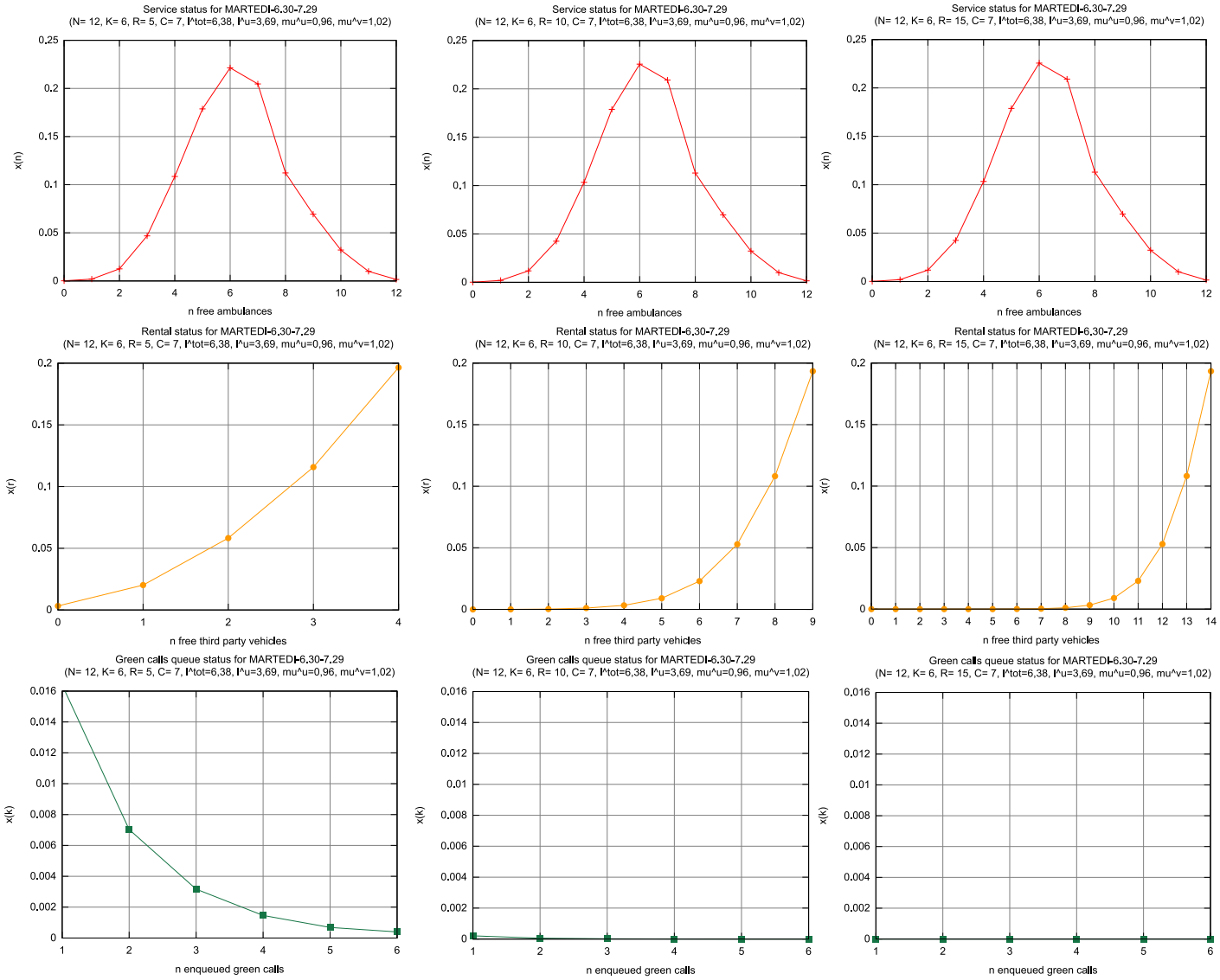


Figura 3.24: grafici del comportamento del sistema combinato al variare della dimensione della flotta di ambulanze di terze parti (parametro R).

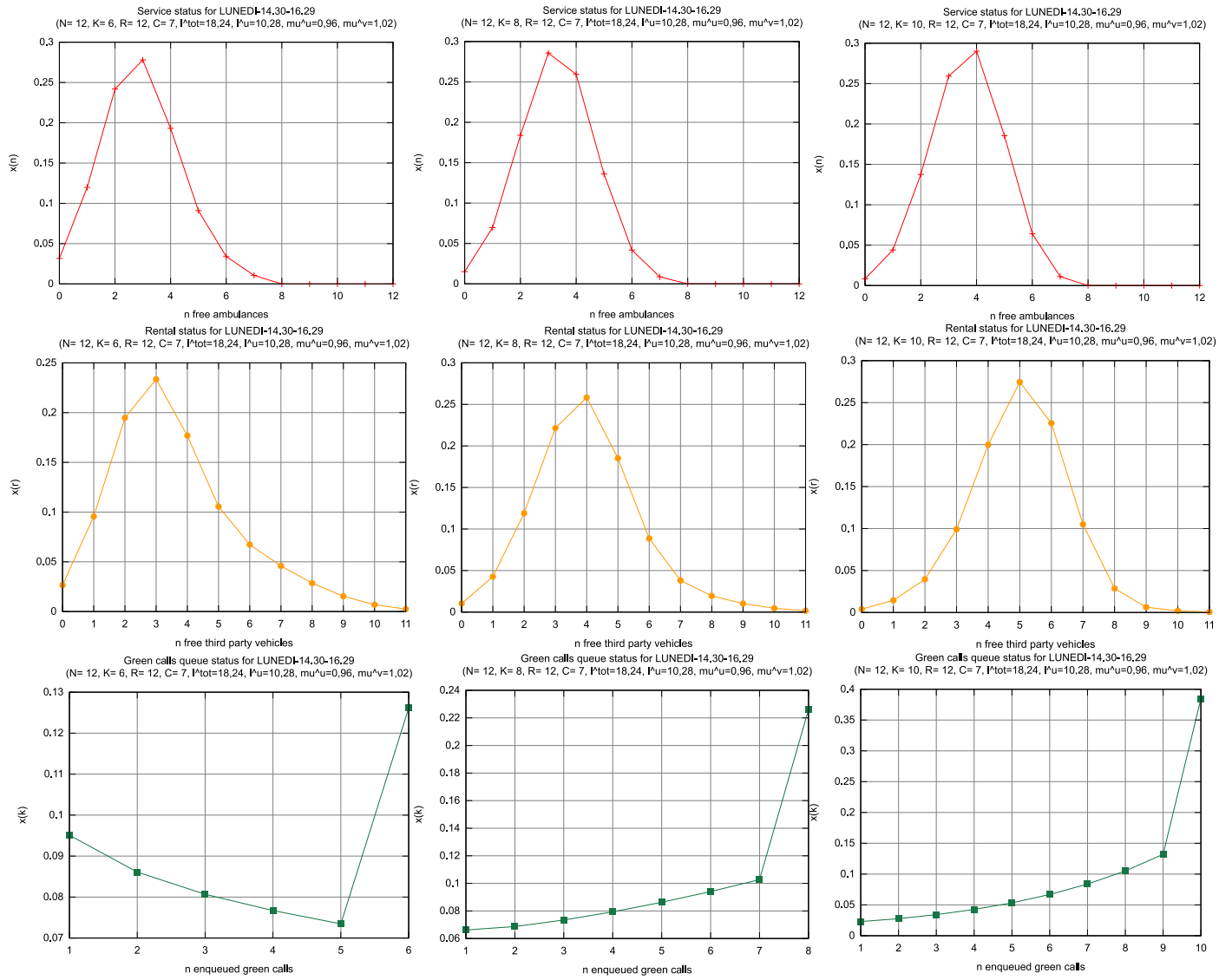


Figura 3.25: grafici del comportamento del sistema combinato al variare della soglia di criticità per le richieste non urgenti (parametro K).

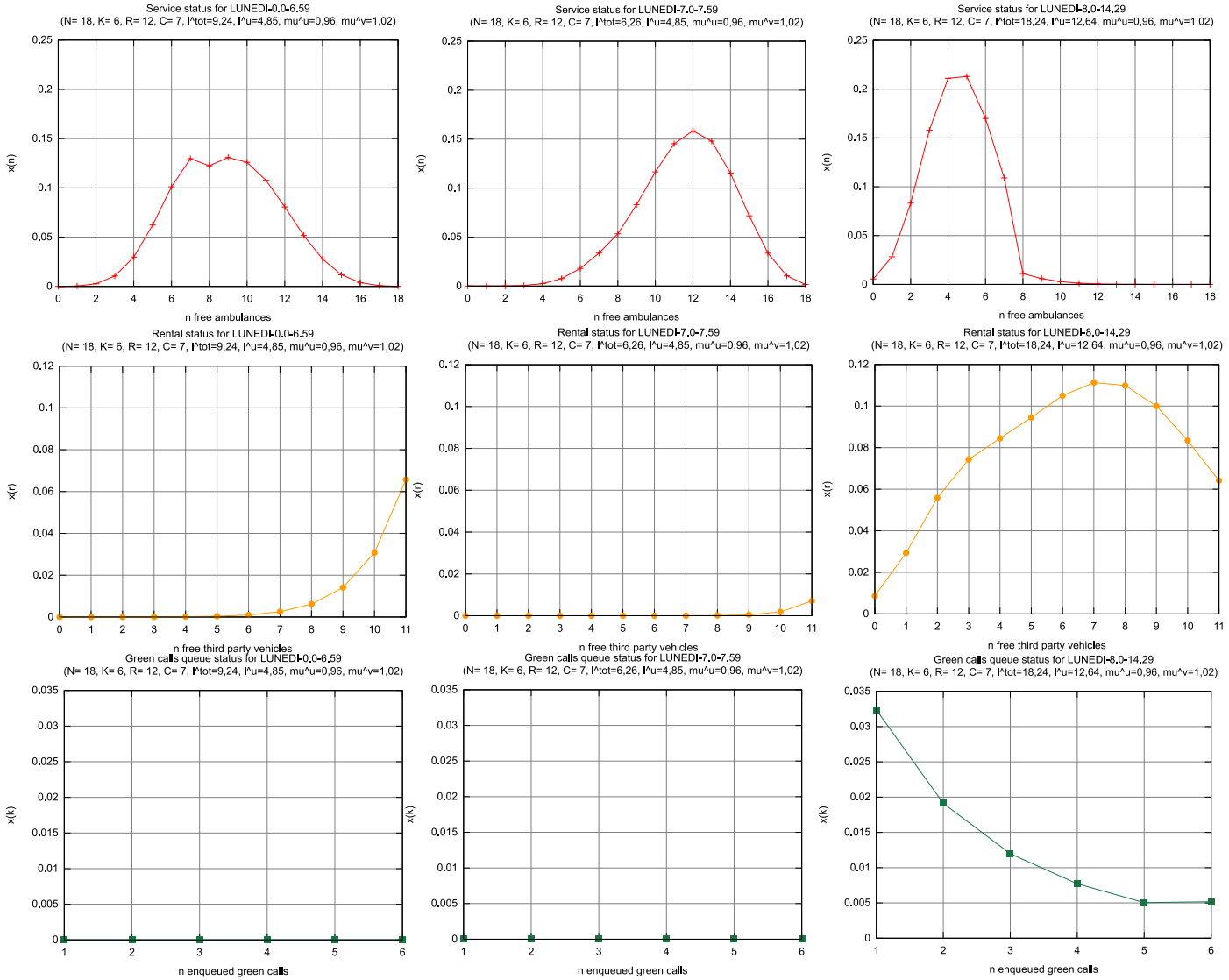


Figura 3.26: grafici del comportamento del sistema combinato al variare della fascia oraria considerata (parametri λ^{tot} e λ^{u}).

3.6 Esempio di dimensionamento del sistema

L'attività di dimensionamento del sistema avviene quotidianamente: il decisore valuta la situazione che, secondo il suo bagaglio di esperienza, potrebbe trovarsi a fronteggiare il giorno successivo. Stimata la possibile criticità della giornata successiva, egli procede a valutare se aumentare o diminuire senza correre rischi le risorse disponibili. A questo proposito si ricordino le considerazioni fatte nell'ambito dell'introduzione al lavoro di tesi (Capitolo 1): spesso il comportamento adottato dal decisore è quello di "andare sul sicuro" puntando sul sovradimensionamento per evitare situazioni rischiose.

Nel seguito del paragrafo si ipotizza un verosimile svolgimento di processo decisionale con il supporto dei modelli visti in questo capitolo. I dati a disposizione sono i risultati sulle analisi di scenario (Capitolo 2) che forniscono una previsione, formulata sulla base dei dati storici, della situazione che ci si troverà ad affrontare; il giorno in questione sarà un tipico mercoledì di Marzo senza nessun evento straordinario. Le fasce orarie risultanti sono così suddivise:

1. 4:00 ÷ 7:59 ($\lambda^{tot} = 7,670$ e $\lambda^{u} = 4,935$), fascia a bassa criticità;
2. 8:00 ÷ 13:59 ($\lambda^{tot} = 25,908$ e $\lambda^{u} = 18,753$), fascia ad altissima criticità dovuta alle attività lavorative e di pendolarismo, è il periodo più difficile di tutta la giornata e sarà il banco di prova per i parametri di dimensionamento;
3. 14:00 ÷ 16:29 ($\lambda^{tot} = 14,475$ e $\lambda^{u} = 12,736$), fascia a media criticità;
4. 16:30 ÷ 20:59 ($\lambda^{tot} = 19,475$ e $\lambda^{u} = 16,743$), ritorno ad un alto livello di criticità per la fascia terminale della giornata.

Per la prima configurazione delle risorse si decide di lasciare invariato l'impiego rispetto alla giornata odierna; utilizzando 12 ambulanze di proprietà, 15 "a gettone" e considerando 6 il limite di guardia per l'accodamento delle richieste non urgenti ($N = 12,\ R = 15,\ K = 6$), si ottiene dal modello combinato (Paragrafo 3.5) il comportamento presentato in Figura 3.27. Come si può notare, nella prima fascia la coda (terzo grafico sulla prima riga) non viene praticamente utilizzata data la bassa incidenza di chiamate di soccorso, così come è molto alta la frazione di tempo in cui la quasi totalità dei mezzi a gettone (secondo grafico) è libera. Nel passaggio alla seconda fascia (seconda riga) la situazione cambia drasticamente: per quasi tutta la sua durata si hanno pochissimi mezzi disponibili con un picco sul valore due sia per

quelle di proprietà (primo grafico) che a noleggio. Come si vede dal grafico di accodamento, la difficoltà della situazione è evidenziata dal fatto che per il 40% del tempo si hanno 6 codici verdi in attesa. La risposta migliora per la terza fascia ma ritorna ad essere di difficile sostenibilità per l'ultima parte della giornata. La situazione che si verrebbe a creare non è accettabile: il decisore prova a migliorarla aumentando il numero di mezzi a noleggio messi in preallarme. Nel secondo caso la flotta di proprietà e la soglia di accodamento rimangono invariati ma vengono mobilitate 5 ambulanze "a gettone" in più ($N = 12,\ R = 20,\ K = 6$). La risposta ottenuta con il modello combinato è presentata in Figura 3.28: come si può notare, il comportamento stimato per la prima fascia non varia rispetto alla prima configurazione; lo scarso carico al quale è sottoposto il sistema non intacca la flotta di mezzi a noleggio il cui utilizzo rimane invariato. Nella seconda fascia, la più critica della giornata, si nota come l'occupazione di ambulanze di proprietà non cambia (il picco di disponibilità si ha sulle due unità) mentre l'incrementato utilizzo di mezzi a gettone permette un miglioramento della situazione della coda: la frazione di tempo nella quale si hanno 6 codici verdi in attesa supera di poco il 10%. Per le ultime due fasce la risposta è sensata: l'utilizzo dei mezzi di proprietà varia di poco a fronte di una migliorata situazione dell'accodamento grazie ad una maggiore disponibilità di mezzi "a gettone". Questa situazione, nonostante sia sostenibile, vede il decisore non pienamente soddisfatto. Si opta quindi per provare a valutare una terza configurazione dove vengono portate a 15 unità le ambulanze di proprietà e si alza la soglia per l'accodamento dei codici verdi ($N = 15$, $R = 20$, $K = 8$). La situazione stimata dal modello è presentata in Figura 3.29; come si può notare dalla fascia ad alta criticità (seconda riga), si ottiene un notevole miglioramento dell'occupazione dei mezzi di proprietà ed una diminuzione dello sfruttamento di quelli a noleggio. Inoltre, nessun livello di accodamento arriva al 10% del tempo della fascia.

Nonostante l'ultima situazione ottenuta sia ampiamente accettabile, il decisore stabilisce che valuterà la bontà della risposta allo scenario anche in base alle frazioni del tempo in cui più di metà della flotta è impegnata in missione. Si veda la Tabella 3.1: con la prima situazione di impiego di risorse (prima riga) per l'82,4% della giornata considerata si ha meno della metà della flotta di proprietà disponibile (colonna PROPRIETÀ); osservando invece la situazione con meno di 7 mezzi liberi, la frazione sale addirittura al 97,2%. Per quanto riguarda i mezzi a noleggio (colonna NOLEGGIO), per il 50% della giornata la loro disponibilità sarà sotto le 6 unità così come si passa al 64% per le 9 unità; l'ultima colonna (CODA) riporta la frazione di tempo della giornata in cui si ha *almeno* un codice verde in coda: per

CONFIGURAZIONE	PROPRIETÀ		NOLEGGIO		CODA
	$n \leq 5$	$n \leq 7$	$r \leq 6$	$r \leq 9$	$k > 0$
Caso 1 $\{N = 12,\ R = 15,\ K = 6\}$	82,4	97,2	50,0	64,7	29,0
Caso 2 $\{N = 12,\ R = 20,\ K = 6\}$	81,8	97,2	26,9	40,0	13,4
Caso 3 $\{N = 15,\ R = 20,\ K = 8\}$	67,1	87,5	18,6	29,9	12,0

Tabella 3.1: frazioni di tempo della fascia oraria nelle quali il sistema viola i vincoli di criticità imposti dal decisore.

il primo caso il valore è pari al 29%. Nel secondo caso, quello relativo al primo intervento correttivo attuato tramite un aumento dei mezzi a noleggio, le frazioni di disponibilità delle proprie ambulanze non variano, così come era stato evidenziato dai grafici: il carico è assorbito dall'incremento di ambulanze di terzi che portano ad avere solo il 13,4% del tempo della giornata in cui si ha almeno un paziente non urgente in coda. Nel caso dell'ultima configurazione si ha finalmente una maggiore disponibilità sia di mezzi propri che di quelli "a gettone" ed il ricorso all'accodamento dei codici verdi scende al 12% del tempo.

Osservando i grafici ed imponendo delle soglie di accettabilità sulle frazioni del tempo della giornata in cui il sistema si trova in stati particolari, il decisore si può avvalere dei risultati dei modelli visti in questo capitolo per dimensionare con maggiore consapevolezza le risorse da dedicare al servizio.

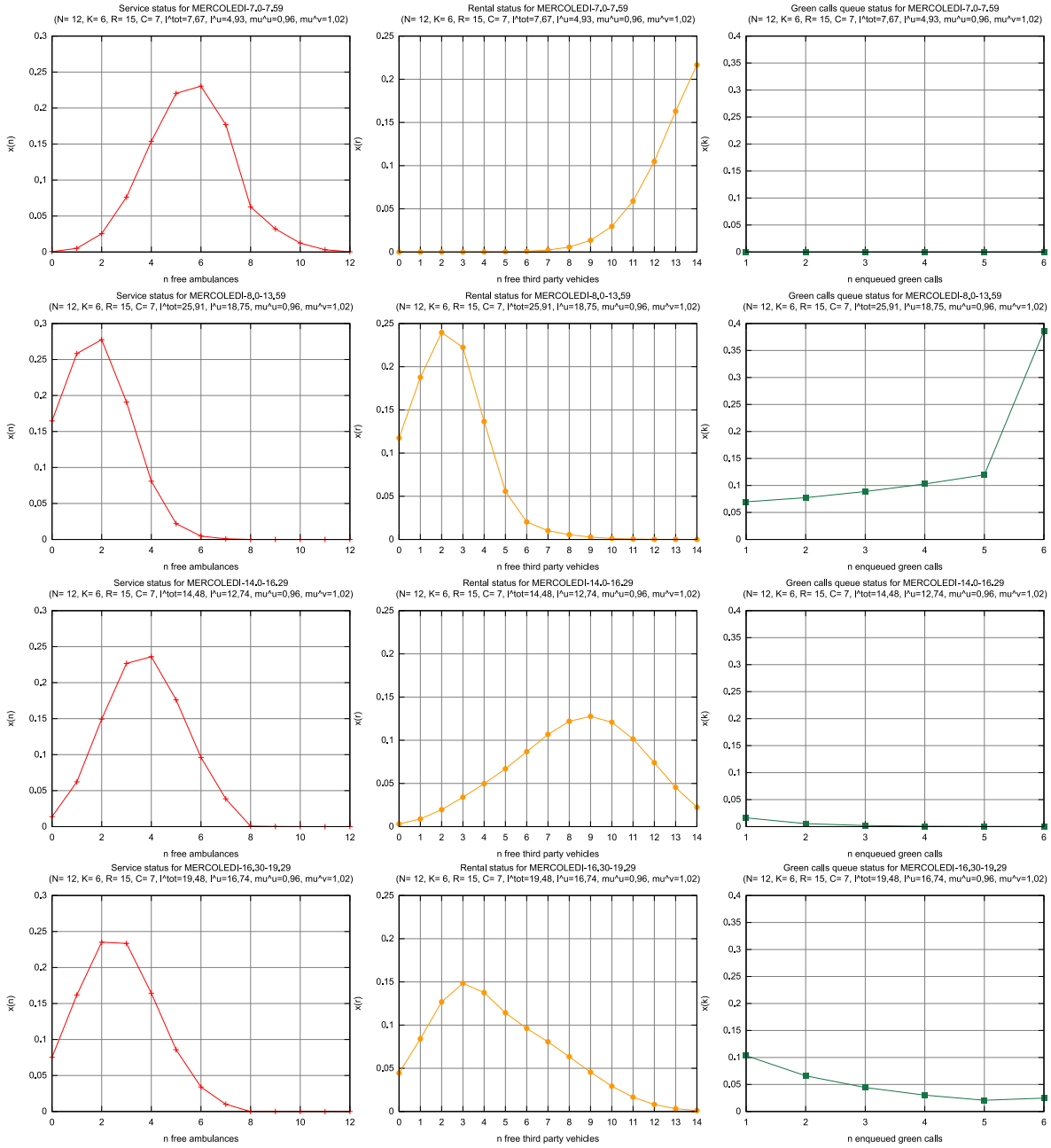


Figura 3.27: processo decisionale di dimensionamento, valutazione della prima configurazione delle risorse ($N = 12$, $R = 15$, $K = 6$).

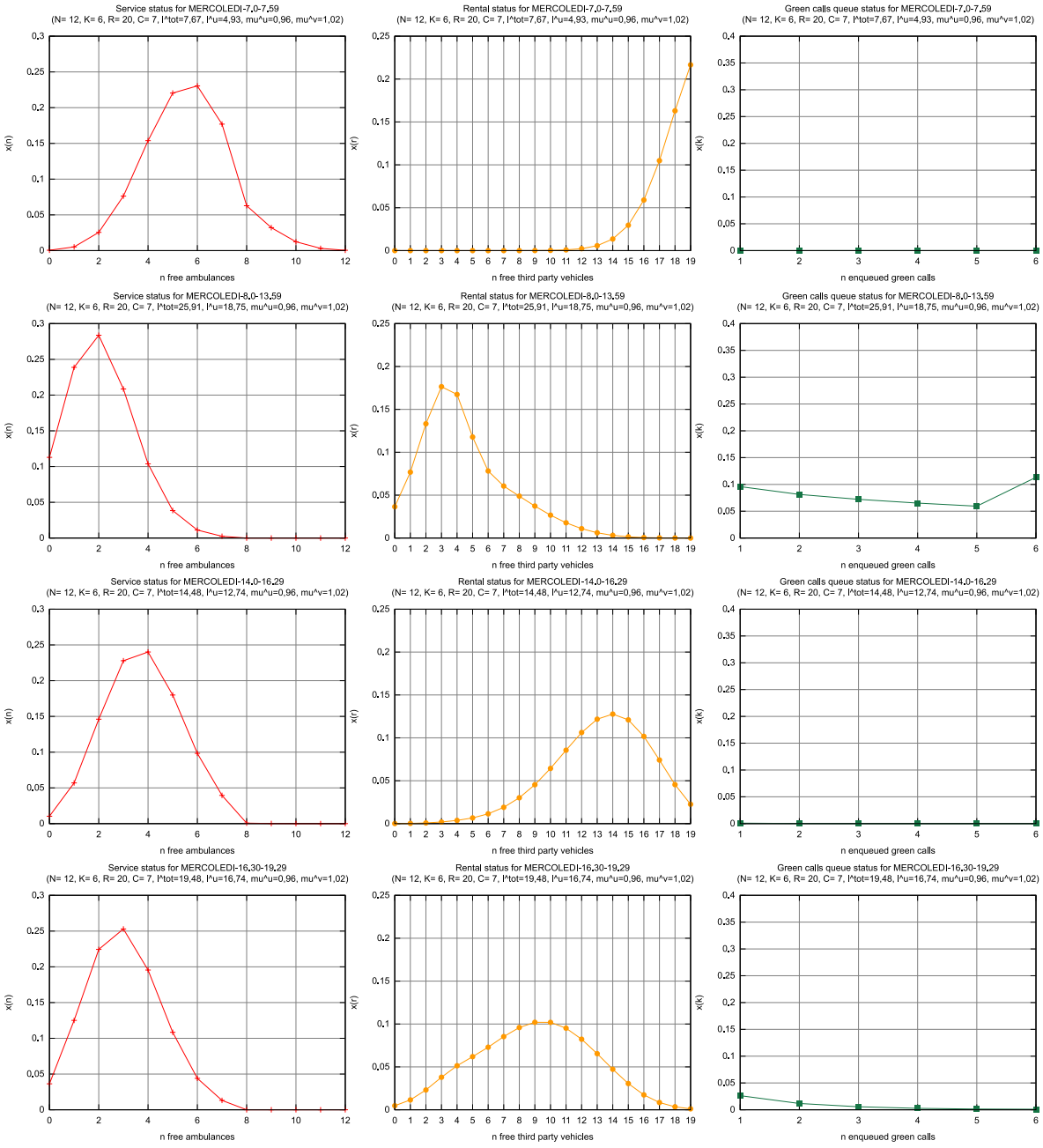


Figura 3.28: processo decisionale di dimensionamento, valutazione della seconda configurazione delle risorse ($N = 12$, $R = 20$, $K = 6$).

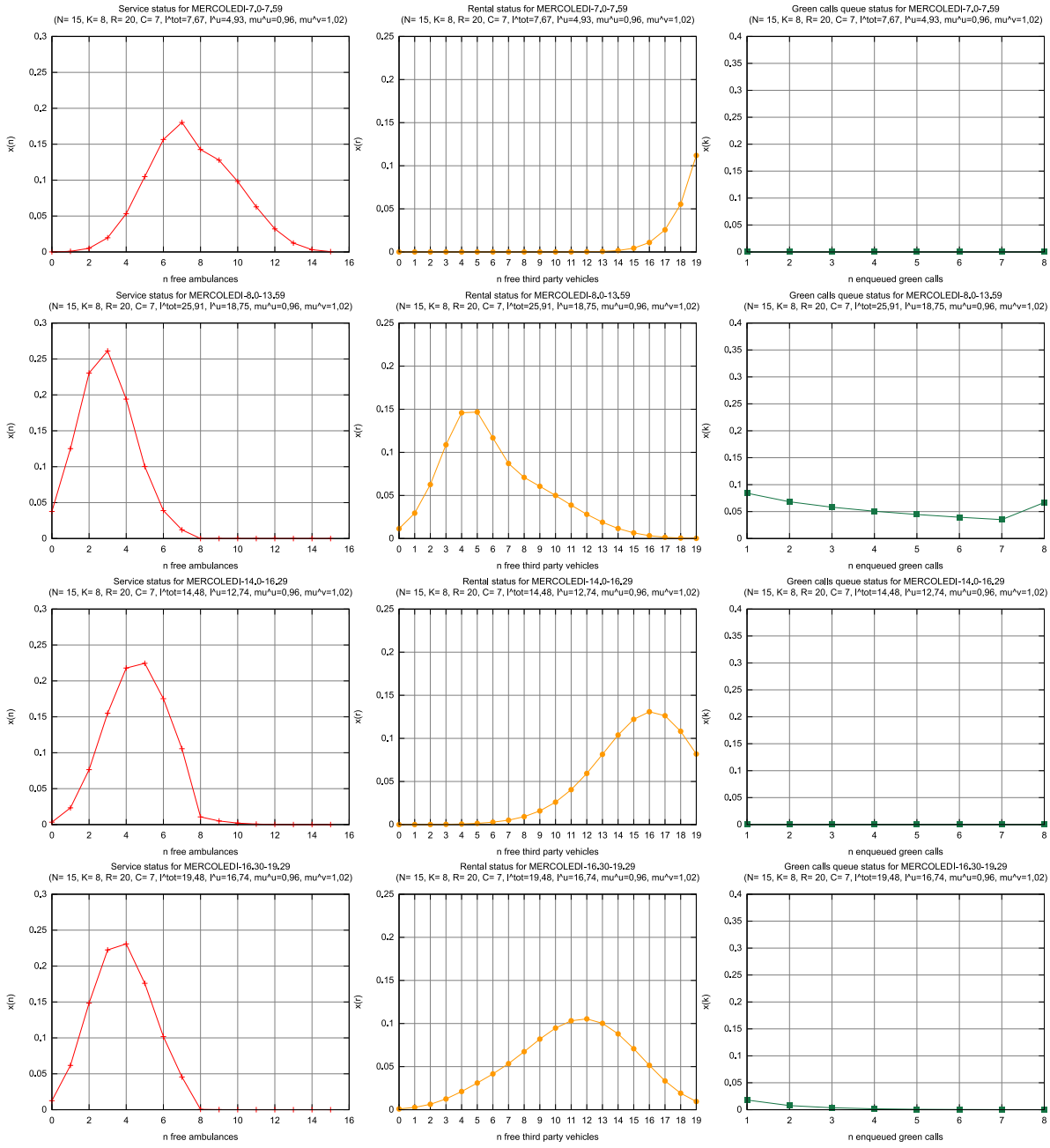


Figura 3.29: processo decisionale di dimensionamento, valutazione della terza e definitiva configurazione delle risorse ($N = 15$, $R = 20$, $K = 8$).

Capitolo 4

Modelli per l’ottimizzazione del servizio sul territorio

Introduzione

I modelli visti nel Capitolo 3 forniscono informazioni utili dal punto di vista strategico, fungendo da supporto alla fase decisionale di dimensionamento del servizio. La “dimensione” di queste informazioni è però unicamente quella temporale: il decisore può interrogare i modelli chiedendo quale sarebbe il comportamento del Servizio “118” in una data fascia oraria a fronte di un dato impiego di risorse. Di fatto, i modelli a coda del Capitolo 3 considerano tutto il territorio come un unico punto senza alcuna conoscenza della sua conformazione. Lo

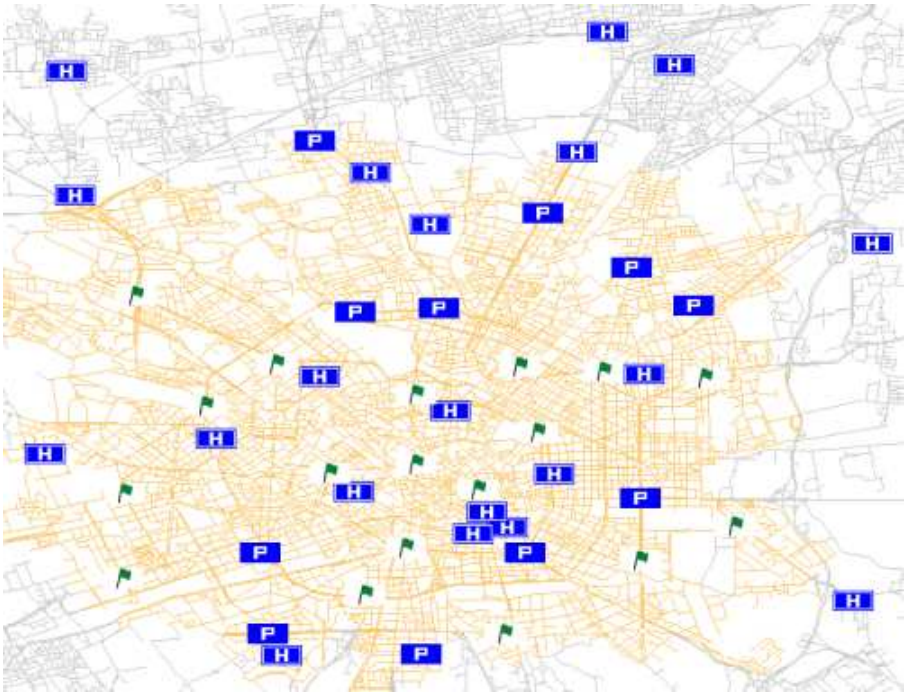


Figura 4.1: esempio di mappa relativa al territorio della città di Milano con configurazione dei mezzi presso i punti d’attesa.

scopo di questa parte del lavoro di tesi è proprio riuscire a fornire ai soggetti incaricati dell’organizzazione del servizio una serie di strumenti che diano informazioni di tipo spaziale. Questi modelli permettono, ad esempio, di determinare per ogni stato del sistema (ovvero quante ambulanze sono disponibili) qual’è il posizionamento che massimizza la copertura del territorio di competenza. Un’altra possibilità è quella di decidere dove costruire i punti d’attesa, scegliendo in un insieme di siti candidati, per fare in modo di ottimizzare la copertura del territorio. In Figura 4.1 è riportato un esempio di risultato prodotto con informazioni

spaziali ottenute dai modelli di ottimizzazione: per ogni mezzo disponibile (rappresentati come bandierine) si è deciso il posizionamento presso le colonnine (rappresentate come “P”) che produce la copertura globalmente ottima del territorio.

L’approccio adottato è quello dell’utilizzo di modelli di programmazione matematica, realizzati per considerare differenti problematiche inerenti all’organizzazione del servizio. La prima versione, presentata nel Paragrafo 4.2, considera il posizionamento ottimo della flotta di ambulanze disponibili presso i punti d’attesa definiti in precedenza. Il secondo modello, trattato nel Paragrafo 4.3, considera sia il posizionamento ottimo dei mezzi, sia un insieme di punti candidati sul territorio fra i quali viene decisa la costruzione di un sottoinsieme ottimo di colonnine d’attesa. Nell’ultimo approccio, trattato nel Paragrafo 4.4, viene massimizzata la copertura complessiva, considerando nel contempo tutti i possibili stati del sistema e limitando le rilocalizzazioni dei mezzi necessarie per passare da una qualsiasi configurazione ottima alla successiva.

Il modello presentato nel secondo paragrafo affronta un problema definito unicamente a livello strategico: la costruzione dei punti di stazionamento può essere valutata solamente in fase di dimensionamento delle infrastrutture del servizio. Il problema della rilocalizzazione dei mezzi invece, nonostante nell’ambito di questo lavoro venga risolto da un punto di vista strategico, ha importanti ripercussioni anche a livello tattico.

4.1 Tassonomia dei modelli presenti in letteratura

Dall’evoluzione dei modelli impiegati in questo ambito, si nota come l’incremento di prestazioni dell’hardware e dei solutori di programmazione matematica abbia portato all’introduzione di tecniche di ricerca operativa in tutti i livelli decisionali (Broncone, La porte e Seme, [3]). A seconda della loro natura, i modelli possono essere classificati come *statici* o *dinamici*: i primi ricavano la localizzazione dei mezzi che produce la copertura ottimale, i secondi sfruttano la possibilità di rilocalizzare continuamente le ambulanze durante la giornata per fornire un migliore servizio al cittadino. È possibile operare un’ulteriore suddivisione in base alla presenza o meno di elementi stocastici: nel caso in cui tutti i dati siano noti si parla di modelli *deterministici*, nel caso contrario in cui si ha la presenza di processi stocastici si tratta di modelli *probabilistici*.

Tra gli esempi presenti in letteratura, il caso più semplice è quello del *location set covering model* o LSCM, presentato da Toregas et al. in [27], il cui obiettivo è la minimizzazione

del numero di ambulanze necessarie a coprire tutti gli utenti sul territorio. Nonostante il LSCM sia un primo approccio ragionevole, nel caso di questo lavoro il modello di posizionamento è definito all'interno di uno scenario con un numero noto a priori di ambulanze, quantità associate agli stati dei modelli a coda visti nel Capitolo 3. Un secondo approccio statico e deterministico è stato presentato da Church e ReVelle in [5]: il *maximal covering location problem*, o MCLP, ha come obiettivo la massimizzazione della popolazione coperta dal servizio a fronte di un vincolo sulla disponibilità di mezzi di soccorso. Questo modello rispecchia esattamente il primo approccio considerato nel Paragrafo 4.2 dove, assegnando un peso ad ogni possibile zona sul territorio, si impone la massimizzazione del valore complessivo di copertura. Entrambi questi esempi di modelli statici e deterministici possono assumere un ruolo significativo a livello *strategico*. Il LSCM può essere utilizzato come strumento di pianificazione delle risorse, durante il dimensionamento della flotta di ambulanze necessarie a servire la totalità dell'area di competenza mentre il secondo, il MCLP, serve fare miglior uso delle limitate risorse a disposizione. Un'applicazione di successo del MCLP è quella realizzata da Eaton et. al in [9] allo scopo di riorganizzare il servizio di emergenza medica della città di Austin in Texas.

Il difetto più evidente, in comune sia al LSCM che al MCLP, è la non considerazione del fatto che, nel momento in cui un mezzo lascia il suo punto di stazionamento per una missione, gli utenti coperti dalla sua presenza smettono di essere serviti. Per ovviare a questo problema sono stati realizzati modelli statici e deterministici con *copertura aggiuntiva*; tutti gli esempi appartenenti a questa categoria sono delle estensioni del MCLP che si basano su di un'assunzione comune secondo la quale, se un utente è coperto da più di una colonnina, qualora uno di questi si svuoti a causa di una missione ce ne sarà un'altra che continuerà a coprirlo. Un primo esempio è quello presentato da Schilling et al. in [24], nato per gestire più tipologie di mezzi differenti. Il *tandem equipment allocation model*, o TEAM, è una diretta estensione del MCLP con l'introduzione di una relazione gerarchica fra le tipologie di mezzi a disposizione. Il TEAM prevede che un utente sia coperto solo se almeno un mezzo di ciascuna tipologia lo può raggiungere entro il tempo limite di soccorso. Un'ulteriore estensione, la *facility-location, equipment-emplacement technique* o FLEET, presentata sempre in [24], prevede la limitazione del numero di siti sui quali è possibile porre in attesa un mezzo. Un approccio più interessante ai fini di questo lavoro è quello proposto da Marianov e ReVelle in [21] dove, nonostante l'appartenenza alla stessa famiglia di TEAM e FLEET, il modello tenta di individuare i siti migliori per costruire delle stazioni di servizio (colonnine nel nostro

caso) ed assicurando che ogni utente sia coperto da almeno una di queste. Nel caso in cui sia coinvolto un unico tipo di veicolo, l'approccio utilizzato è quello di una modifica diretta del modello MCLP per fornire una migliore copertura multipla senza tuttavia modificare la dimensione della flotta. Come suggerito da Daskin e Stern ([6]) e da Hogan e ReVelle ([17]), una seconda funzione obiettivo può essere aggiunta al MCLP per operare una distinzione sulle sue multiple soluzioni ottime. Nel primo caso, il modello utilizza una seconda funzione obiettivo che massimizza il numero di utenti coperti da più di un punto d'attesa, nel secondo caso viene massimizzato il valore di copertura globale che risulta come contributo di almeno due mezzi differenti. Questo concetto di copertura multipla, detto *copertura di riserva* (*backup coverage*), è incorporato anche nei modelli BACOP1 e BACOP2 presentati in [17]: in questi approcci vengono utilizzati due diversi livelli di copertura, il primo dove un utente è servito se un'ambulanza è posizionata entro distanza di copertura, il secondo dove un utente è considerato servito solo se sono due le ambulanze entro distanza massima. È possibile pesare il contributo di ognuno dei due nella funzione obiettivo. Il *double standard model*, o DSM, proposto da Gendreau, Laporte e Semet ([12]) prevede l'utilizzo di due differenti distanze massime di copertura, una strettamente minore della seconda. Il modello impone che tutti gli utenti abbiano almeno una colonnina entro la maggiore delle distanze mentre una frazione della domanda totale deve essere coperta entro la minore; l'obiettivo massimizza il numero di utenti coperti due volte entro la distanza più stringente.

Per tentare di considerare la natura stocastica delle richieste di emergenza, sono stati sviluppati modelli statici e probabilistici con copertura aggiuntiva. Un primo approccio, proposto da Daskin in [7], ha portato alla realizzazione del *maximum expected covering location problem* o MEXCLP. In questa formulazione viene introdotto il concetto di *frazione di occupazione* (*busy fraction*), ovvero la probabilità che un'ambulanza non riesca a servire una richiesta in quanto già occupata in missione, indicata dall'autore come il rapporto fra il valore atteso della durata di una chiamata ed il numero totale di mezzi disponibili. Il MEXCLP impone la massimizzazione della copertura considerando però che non sempre un mezzo è immediatamente pronto a partire, con una certa probabilità sarà impegnato altrove. Fujiwara et al. hanno presentato in [11] un'applicazione del MEXCLP alla città di Bangkok. Una prima estensione al MEXCLP, chiamata TIMEXCLP, è stata definita da Repede e Bernardo ([22]) e prevede esplicitamente la possibilità che i tempi di percorrenza dei mezzi varino durante la giornata; il modello è stato combinato con un modulo di simulazione per validare i risultati. In modo molto simile al TIMEXCLP, Goldberg et al. ([14]) hanno definito un modello in cui

i tempi di percorrenza sono un processo stocastico e la funzione obiettivo impone la massimizzazione del valore atteso degli utenti serviti entro il tempo limite. ReVelle ed Hogan ([23]) propongono due versioni del loro *maximum availability location problem* (MALP I e MALP II) dove viene imposta la massimizzazione degli utenti che risultano coperti con un livello di probabilità dato. Il primo, il MALP I, considera la frazione di occupazione identica per tutti i mezzi e massimizza il numero di utenti che sono serviti da una quantità minima di ambulanze con una probabilità superiore alla soglia data. Nel MALP II invece, l'assunzione che la frazione di occupazione sia uguale per tutti i mezzi viene rilassata. Questo modello richiede quindi un valore di probabilità di occupazione di una certa ambulanza o con quale probabilità ci si troverà ad avere un dato numero di mezzi occupati. Questi valori sono stati calcolati utilizzando un approccio analitico come il *modello ad ipercubo* di Larson ([19]), del tutto simile a quanto visto per i modelli a coda del Capitolo 3. Data la disponibilità dello strumento analitico sviluppato dal Larson, Batta et al. ([1]) hanno esteso il MEXCLP creando l'AMEXCLP (*Adjusted* MEXCLP) dove, nella funzione obiettivo, ogni termine è pesato da un fattore che tiene conto che le ambulanze non operano in modo totalmente indipendente ma possono essere viste come i serventi in un sistema a coda. Da questa assunzione viene naturale l'utilizzo dell'ipercubo di Larson per il calcolo delle frazioni di occupazione.

Nella localizzazione dei mezzi di soccorso, è possibile prevedere decisioni di rilocalizzare le ambulanze dinamicamente con lo scopo di non lasciare aree scoperte. Questo comporta la necessità di ricalcolare continuamente, in occasione di ogni chiamata, la strategia di rilocalizzazione ottima basata sulle informazioni a disposizione. In questo ambito, l'unico modello esistente è stato realizzato da Gendreau et al. ([13]) ed è basato sull'estensione del DSM. In aggiunta a quanto visto per quest'ultimo, il modello *dinamico* tiene conto di alcune considerazioni di carattere pratico: rilocalizzazioni successive non possono essere le stesse, viaggi ripetuti fra le stesse colonnine devono essere evitati così come spostamenti troppo lunghi. Il *dynamic double standard model at time t* o DDSM^t, è risolto ogni qualvolta si presenta una chiamata e la funzione obiettivo è identica a quella vista per il DSM, penalizzata però dai costi delle rilocalizzazioni. La messa in opera di questo modello ha portato gli autori a realizzare un'euristica basata sul tabu search perennemente in esecuzione parallela all'interno di un cluster.

4.2 Ottimizzazione della copertura

Il problema più immediato, dal carattere spiccatamente "tattico", che il decisore si trova ad affrontare è quello di servire al meglio il territorio entro il quale si trova ad operare. L'obiettivo che deve perseguire è quello di raggiungere il maggior numero di clienti, o di zone, con una copertura che assicuri un servizio accettabile: ciò si traduce nel tentativo di posizionare i mezzi a disposizione, facendo in modo che il maggior numero di clienti sia raggiungibile in un tempo inferiore al limite di servizio per le urgenze. Questo obiettivo è ovviamente raggiunto in una situazione di inattività: si presuppone che, quando tutta la flotta è disponibile, ogni zona sia servita entro il tempo limite. La realtà pone però il servizio in situazioni non ideali; come si è visto grazie alle informazioni prodotte dai modelli strategici a coda (Capitolo 3), le frazioni del tempo in cui *tutti* i mezzi sono disponibili sono una porzione infinitesima rispetto a stati enormemente più probabili. Questo fa sì che, data la dimensione della flotta di ambulanze, per ogni possibile numero di mezzi liberi (da uno solo alla totalità della flotta) esista una configurazione di dislocazione presso le colonnine che massimizzi la copertura del territorio. Grazie a questo primo modello di programmazione lineare intera è possibile, dato un qualsiasi numero di ambulanze libere, ricavare questa configurazione.

I parametri su cui si basa il modello di ottimizzazione sono:

- $N \in \mathbb{N}^+$ numero di utenti potenziali. Con il termine *utente potenziale* si fa riferimento ad un qualsiasi punto del grafo (prodotto nell'ambito del Capitolo 2) dal quale è possibile che si generi almeno una domanda di servizio.

- $w_n \in \mathbb{N}$ con $n = 1, \ldots, N$, vettore dei pesi associati ad ogni utente potenziale. È necessario assegnare un peso ad ogni nodo del grafo per considerare il numero di richieste che potrà generare. Considerando quanto detto nel Capitolo 2, questo dato viene ricavato assegnando ogni richiesta di missione presente nello storico ad un nodo del grafo.

- $C \in \mathbb{N}^+$ numero di punti di stazionamento sul territorio.

- $V \in \mathbb{N}^+$ indica il numero complessivo di mezzi disponibili. In questa quantità sono comprese anche le ambulanze "a gettone", senza distinzione.

- $\overline{d} \in \mathbb{N}^+$ distanza massima di copertura, espressa in chilometri. Questo dato rappre-

senta la distanza che un mezzo riesce a coprire, in una data fascia oraria, entro il tempo massimo di servizio per i codici gialli e rossi (pari ad 8 minuti).

- $d_{c,n} \in \mathbb{N}^2$ con $n = 1, \ldots, N$ e $c = 1, \ldots, C$, matrice delle distanze che intercorrono fra ogni colonnina c e qualsiasi utente potenziale n.

Le variabili considerate da questo modello di copertura sono:

- $y_c \in \{0, 1\}$ con $c = 1, \ldots, C$. Vettore di variabili binarie che rappresentano la presenza di un'ambulanza in ognuna delle C colonnine. Se y_c ha valore uno, un mezzo è stato allocato presso il punto d'attesa c, in caso contrario il valore è zero.

- $z_n \in \{0, 1\}$ con $n = 1, \ldots, N$. Vettore di variabili binarie che segnalano se ognuno degli N clienti potenziali è coperto dal servizio. Per essere servito, il punto n deve avere *almeno* un mezzo posizionato presso una colonnina entro la distanza massima di servizio $\overline{d}$. Se questo accade, z_n ha valore uno, altrimenti ha valore zero.

Per l'ottimizzazione della copertura del territorio è stato definito il seguente modello:

$$\max \quad \sum_{n=1}^{N} w_n \, z_n \tag{4.1}$$

$$s.t. \quad \sum_{c=1}^{C} y_c = V \tag{4.2}$$

$$z_n \leq \sum_{c=1,\ldots,C \,:\, d_{c,n} \leq \overline{d}} y_c \qquad \forall \, n = 1, \ldots, N \tag{4.3}$$

$$y_c \in \{0, 1\} \qquad \forall \, c = 1, \ldots, C$$

$$z_n \in \{0, 1\} \qquad \forall \, n = 1, \ldots, N$$

Come si può notare dalla funzione obiettivo (4.1), viene massimizzata la copertura complessiva degli utenti potenziali, pesati secondo la loro reale importanza. Per fare ciò, l'argomento della massimizzazione è la somma su tutti gli utenti del prodotto fra la variabile binaria di copertura z_n ed il peso w_n associato ad ognuno.

Il primo vincolo al quale è soggetto il modello, il (4.2), riguarda la dimensione della flotta. Esso infatti limita la quantità di ambulanze dislocate presso i punti d'attesa al numero di mezzi effettivamente disponibili.

L'ultimo vincolo (4.3), impone che ognuna delle variabili binarie z_n abbia valore minore o uguale alla somma su tutte le colonnine delle variabili di occupazione da parte di un'ambulanza. Si noti che la sommatoria è condizionata: vengono infatti considerate solo le colonnine tali per cui la distanza fra la colonnina e l'utente potenziale $d_{c,n}$ è minore o uguale alla distanza massima di copertura $\overline{d}$.

Risultati

Il modello è stato scritto in linguaggio *GNU MathProg* ([15]) ed è stato utilizzato CPLEX (versione 8.11) come solutore. I test sono stati condotti al fine di valutare la possibilità di trattare istanze di dimensioni crescenti [1]; a questo proposito si è deciso di analizzare l'andamento dei tempi spesi da CPLEX al variare del numero di colonnine (parametro C), l'unica variabile sotto il diretto controllo del decisore. Per i test è stato utilizzato il grafo completo relativo alla città di Milano composto da $N = 12442$ nodi, ognuno considerato come un utente potenziale. Il numero di ambulanze utilizzato è $V = 18$ e la distanza limite per la copertura è pari a $\overline{d} = 3,333$ km, corrispondente a circa 25 km/h di velocità media di percorrenza.

In Tabella 4.1 è riportato l'andamento dei tempi di calcolo di CPLEX (colonna T_SOLVE) al variare del numero di colonnine (colonna C). Dai valori si può osservare come il problema sia di facile risoluzione anche per istanze di notevoli dimensioni: persino nel caso irrealistico di 1000 punti di stazionamento, il tempo impiegato per trovare la soluzione ottima è di poco superiore al secondo (si consideri che in una situazione dalle dimensioni nell'ordine di Milano, i punti d'attesa superano raramente le 50 unità). L'andamento dei tempi è rappresentato in Figura 4.2.

[1] Tutti i test di questo capitolo sono stati effettuati utilizzando un PC con CPU Intel Dual Core a 2 GHz e 2 GB di RAM, CPLEX versione 8.11 ed interprete AMPL su S.O. GNU/Linux 2.6.22.

C	T_SOLVE
30	$0,03$
40	$0,05$
50	$0,05$
100	$0,07$
150	$0,09$
200	$0,11$
250	$0,19$
300	$0,26$
350	$0,35$
400	$0,49$
450	$0,63$
500	$0,72$
550	$0,82$
600	$0,85$
800	$0,90$
1000	$1,03$

Tabella 4.1: andamento dei tempi di calcolo al variare del numero di punti candidati.

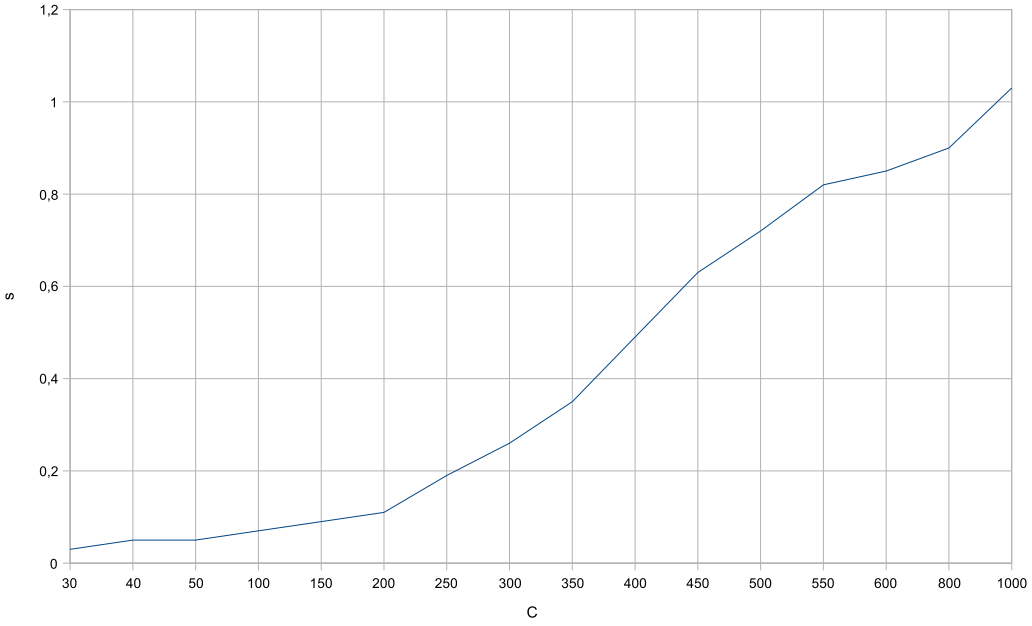


Figura 4.2: andamento dei tempi di calcolo al variare del numero di punti d'attesa disponibili.

4.3 Ottimizzazione della copertura con costruzione dei punti di stazionamento

Un secondo problema al quale il decisore deve far fronte è quello del posizionamento dei punti d'attesa. Sul territorio viene identificato un insieme di *punti candidati*, luoghi adatti ad accogliere un mezzo durante lo stazionamento; le colonnine possono essere realmente costruite solo in uno qualsiasi di questi punti. L'obiettivo del decisore rimane quello di coprire il maggior numero di utenti potenziali: in questo caso però, la copertura viene considerata complessivamente lungo una serie di fasce orarie. L'ottimizzazione su più fasce contemporaneamente è dettata dal fatto che, dovendo decidere la costruzione dei punti d'attesa ed essendo questa una decisione definitiva, se anche una sola fascia dovesse richiedere la presenza di una colonnina presso un punto candidato allora la costruzione sarebbe necessaria. La funzione svolta da questo modello è decidere la costruzione delle colonnine presso i punti candidati posizionandovi i mezzi in attesa in ognuna delle fasce orarie specificate. Nel fare questo, l'obiettivo complessivo rimane sempre la massimizzazione della copertura del servizio.

I parametri su cui si basa il modello di ottimizzazione sono:

- $N \in \mathbb{N}^+$ numero di utenti potenziali. Il termine *utente potenziale* assume il significato spiegato nel paragrafo precedente.

- $T \in \mathbb{N}^+$ numero di tipologie di fasce temporali da considerare. Il modello richiede di ottimizzare la copertura lungo tutte le tipologie di fasce temporali specificate dal decisore (ad esempio, lungo tutti i lunedì mattina ed i martedì pomeriggio, ecc...).

- $w_{n,t} \in \mathbb{N}$ con $n = 1, \ldots, N$ e $t = 1, \ldots, T$, vettore dei pesi associati ad ogni utente potenziale in ognuna delle fasce temporali. In questo caso il dato deve essere necessariamente collocato temporalmente: il singolo peso $w_{n,t}$ viene fissato contando il numero di richieste di emergenza che hanno avuto effettivamente origine nel nodo n durante la fascia di tipo t.

- $B \in \mathbb{N}^+$ numero di *punti candidati* identificati.

- $\overline{B} \in \mathbb{N}^+$ capacità di costruzione. Questo valore rappresenta il numero massimo di colonnine che il decisore è effettivamente in grado di realizzare.

- $V \in \mathbb{N}^+$ indica il numero complessivo di mezzi disponibili indipendentemente dalla natura del contratto al quale sono sottoposte le condizioni di utilizzo. Per fare in modo che il problema abbia senso e dato che in ogni colonnina può sostare un solo mezzo, dev'essere $V < \overline{B}$.

- $\overline{d}_t \in \mathbb{N}^+$ con $t = 1, \ldots, T$, distanza massima di copertura espressa in metri per la fascia di tipo t.

- $d_{b,n} \in \mathbb{N}^2$ con $b = 1, \ldots, B$ e $n = 1, \ldots, N$, matrice delle distanze che intercorrono fra ogni punto candidato b e qualsiasi utente potenziale n.

Le variabili considerate da questo modello sono:

- $y_{b,t} \in \{0, 1\}$ con $b = 1, \ldots, B$ e $t = 1, \ldots, T$. Matrice di variabili binarie che rappresentano la presenza di un'ambulanza in ognuna delle B potenziali colonnine.

- $z_{n,t} \in \{0, 1\}$ con $n = 1, \ldots, N$ e $t = 1, \ldots, T$. Matrice di variabili binarie che segnalano se ognuno degli N clienti potenziali è coperto dal servizio durante ognuna delle tipologie di fasce orarie considerate.

- $x_b \in \{0, 1\}$ con $b \in 1, \ldots, B$. Vettore di variabili binarie che segnalano se sia necessario o meno costruire una colonnina presso ognuno dei punti candidati.

Per l'ottimizzazione della copertura con costruzione dei punti d'attesa è stato definito il seguente modello:

$$\max \quad \sum_{n=1}^{N} \sum_{t=1}^{T} w_{n,t}\, z_{n,t} \tag{4.4}$$

$$s.t. \quad \sum_{b=1}^{B} y_{b,t} = V \qquad \forall\, t = 1, \ldots, T \tag{4.5}$$

$$z_{n,t} \leq \sum_{b=1,\ldots,B\,:\, d_{b,n} \leq \overline{d}_t} y_{b,t} \qquad \forall\, n = 1, \ldots, N;\ t = 1, \ldots, T \tag{4.6}$$

$$y_{b,t} \leq x_b \qquad \forall\, b = 1, \ldots, B;\ t = 1, \ldots, T \tag{4.7}$$

$$\sum_{b=1}^{B} x_b = \overline{B} \tag{4.8}$$

$$y_{b,t} \in \{0, 1\} \qquad \forall\, b = 1, \ldots, B;\ t = 1, \ldots, T$$

$$z_{n,t} \in \{0,1\} \qquad \forall\, n = 1,\ldots,N;\ t = 1,\ldots,T$$
$$x_b \in \{0,1\} \qquad \forall\, b = 1,\ldots,B$$

Anche in questo caso, la funzione obiettivo (4.4) massimizza il servizio agli utenti, ovvero il prodotto fra il valore della variabile binaria di copertura $z_{n,t}$ ed il peso $w_{n,t}$ assegnato ad ognuno degli utenti potenziali. Come si può notare però, in questa variante del modello di programmazione matematica ciò che si massimizza è il valore di copertura lungo tutte le fasce orarie specificate: è per questo motivo che, oltre alla sommatoria su tutti gli N utenti potenziali, è necessaria anche quella su tutte le tipologie T di fasce.

Il primo vincolo, il (4.5), limita il numero dei mezzi allocabili. Per ognuna delle T fasce infatti, la somma su tutti i B punti candidati delle variabili binarie $y_{b,t}$ di posizionamento dei mezzi deve rispettare il vincolo sul numero massimo di ambulanze disponibili. Dato che la sommatoria *deve* avere valore V, questo vincolo obbliga anche ad usare tutti i mezzi disponibili.

Il secondo vincolo, il (4.6), limita il valore delle variabili di copertura degli utenti. Per ognuna delle T fasce orarie e per ognuno degli N utenti potenziali, il valore della variabile $z_{n,t}$ deve avere valore minore o uguale alla somma, su tutti i B siti candidati, della variabile $y_{b,t}$. Se $z_{n,t}$ può assumere valore uno, significa che nella fascia t c'è *almeno* un punto candidato entro la distanza limite presso il quale è stata assegnata un'ambulanza. Si noti che la somma è condizionata: vengono considerati solamente quei punti candidati che distano al massimo $\overline{d}_t$, la massima distanza che è possibile coprire durante la fascia oraria t entro il tempo limite (8 minuti).

Il terzo vincolo, il (4.7), controlla il valore delle variabili di costruzione dei punti d'attesa. Il controllo viene effettuato per ognuno dei B punti candidati: se il sito b ospita un'ambulanza in *almeno* una delle T fasce orarie, allora è necessario che venga costruita una colonnina.

L'ultimo vincolo, il (4.8), controlla le risorse disponibili per la costruzione dei punti d'attesa. Quest'ultimo limita infatti la somma effettuata su tutti i B punti candidati, delle variabili di costruzione x_b: la quantità complessiva di queste ultime che possono avere valore uno è $\overline{B}$. In questo modo si assicura che il numero di colonnine costruite rispetti l'effettiva capacità di realizzazione. La risoluzione del problema posiziona quindi le ambulanze disponibili presso i punti candidati individuati sul territorio al fine di massimizzare il numero di chiamate di emergenza raggiungibili entro il tempo limite. Il posizionamento comporta la costruzione di una colonnina presso il luogo utilizzato per lo stazionamento: la realizzazione delle

colonnine è vincolata alle risorse disponibili.

La tecnica di soluzione adottata

L'introduzione della problematica relativa alla costruzione dei punti d'attesa ha reso estremamente più difficile la risoluzione del problema rispetto a quanto avviene per il modello visto nel Paragrafo 4.2. Quest'ultimo infatti è un classico problema di *set covering* mentre quanto visto in questo caso ha introdotto una complicazione aggiuntiva.

La natura dei vincoli permette però una scomposizione in due sotto-problemi di facile soluzione se considerati singolarmente. Il vincolo (4.7) è l'unico che mette in relazione le variabili di costruzione x_b e quelle di posizionamento dei mezzi $y_{b,t}$. È possibile rilassando quest'ultimo, separare i due sotto-problemi riguardanti *solamente* la costruzione delle colonnine *o* il posizionamento dei mezzi: sfruttando il *rilassamento lagrangeano* (Beasley, [2]) del vincolo (4.7), la *lagrangeana* ottenuta con la funzione obiettivo (4.4) diventa:

$$\mathcal{L} \;=\; \sum_{n=1}^{N}\sum_{t=1}^{T} w_{n,t}\, z_{n,t} \;-\; \sum_{b=1}^{B}\sum_{t=1}^{T} \mu_{b,t}\,(y_{b,t} - x_b) \tag{4.9}$$
$$\mu_{b,t} \geq 0 \qquad \forall\, b \in 1,\ldots,B;\; t \in 1,\ldots,T$$

Si noti la necessaria introduzione dei *moltiplicatori lagrangeani* (le variabili $\mu_{b,t}$) che, essendo riferiti ad un vincolo di disuguaglianza, sono quantità reali non negative.

Eliminando il vincolo rilassato (4.7) dalla formulazione originaria e considerando come obiettivo la massimizzazione della lagrangeana (4.9) si ottiene il problema seguente:

$$\max \quad \sum_{n=1}^{N}\sum_{t=1}^{T} w_{n,t}\, z_{n,t} \;-\; \sum_{b=1}^{B}\sum_{t=1}^{T} \mu_{b,t}\,(y_{b,t} - x_b)$$

$$s.t. \quad \sum_{b=1}^{B} y_{b,t} = V \qquad \forall\, t = 1,\ldots,T$$

$$z_{n,t} \leq \sum_{b=1,\ldots,B\,:\,d_{b,n}\leq \overline{d}_t} y_{b,t} \qquad \forall\, n = 1,\ldots,N;\; t = 1,\ldots,T$$

$$\sum_{b=1}^{B} x_b = \overline{B}$$

$$y_{b,t} \in \{0,1\} \qquad \forall\, b = 1,\ldots,B;\; t = 1,\ldots,T$$

$$z_{n,t} \in \{0,1\} \qquad \forall\, n = 1,\dots,N;\ t = 1,\dots,T$$
$$x_b \in \{0,1\} \qquad \forall\, b = 1,\dots,B$$

che corrisponde alla formulazione del rilassamento.

Osservando il modello ottenuto si può notare come sia possibile separare in due insiemi disgiunti le variabili dal momento che l'unico vincolo che li legava è stato eliminato. Tramite la separazione delle variabili sono stati ricavati due sotto-problemi; il primo riguarda solo le variabili $y_{b,t}$ e $z_{n,t}$ e la sua formulazione è ottenuta dai soli vincoli che riguardano queste due variabili. Inoltre è stato possibile scomporre ulteriormente questo primo sotto-problema, impostandone uno indipendente per ogni tipologia di fascia oraria considerata. Tutto il modello diventa quindi funzione di t: se lo si indica con $\mathcal{L}_1(t)$, esso va ripetuto per ogni $t = 1,\dots,T$:

$$\begin{aligned}
\max \quad & \sum_{n=1}^{N} w_{n,t}\, z_{n,t} \;-\; \sum_{b=1}^{B} \mu_{b,t}\, y_{b,t} \\
s.t. \quad & \sum_{b=1}^{B} y_{b,t} = V \\
& z_{n,t} \le \sum_{b=1,\dots,B\,:\, d_{b,n} \le \overline{d}_t} y_{b,t} && \forall\, n = 1,\dots,N \\
& y_{b,t} \in \{0,1\} && \forall\, b = 1,\dots,B \\
& z_{n,t} \in \{0,1\} && \forall\, n = 1,\dots,N
\end{aligned}$$

La funzione obiettivo è data dai termini che, nell'ambito della somma della (4.9), coinvolgono $y_{b,t}$ e $z_{n,t}$.

Il secondo sotto-problema, indicato con $\mathcal{L}_2$, riguarda i vincoli e la parte di lagrangeana che coinvolgono le variabili di costruzione x_b. La formulazione del rilassamento risulta:

$$\begin{aligned}
\max \quad & \sum_{b=1}^{B} \sum_{t=1}^{T} \mu_{b,t}\, x_b \\
s.t. \quad & \sum_{b=1}^{B} x_b = \overline{B} \\
& x_b \in \{0,1\} && \forall\, b = 1,\dots,B
\end{aligned}$$

dove l'obiettivo diventa la massimizzazione del valore dei termini che coinvolgono x_b del-

la (4.9); l'unico vincolo rimasto è quello relativo al limite di costruzione (4.8).

Per ottenere il valore della lagrangeana è necessario quindi risolvere $\mathcal{L}_2$ e T volte $\mathcal{L}_1(t)$ per poi calcolarne la somma:

$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_1(t) + \mathcal{L}_2 \tag{4.10}$$

Il valore risultante dalla (4.10) si riferisce ad una soluzione che non è necessariamente ammissibile per la formulazione iniziale del problema in quanto manca del vincolo rilassato che, di conseguenza, potrebbe non essere rispettato. Per riportare all'ammissibilità la soluzione ottenuta dal rilassamento è stata realizzata un'*euristica lagrangeana* che, partendo da una soluzione non ammissibile, la modifica per ottenere una soluzione rispettosa del vincolo (4.7).

Algoritmo 3 : euristica lagrangeana per il calcolo del lower bound per l'ottimizzazione della copertura con costruzione dei punti di stazionamento.

$u_b \in \mathbb{N}$ con $b = 1, \ldots, B$
$B^* = \varnothing$
▷ calcolo utilizzo punti candidati
for each $b = 1, \ldots, B$ **do**
    $u_b = \sum_{t=1}^{T} y_{b,t}$
end for
▷ individuazione punti più sfruttati
while $|B^*| < \overline{B}$ **do**
    $B^* = B^* \cup \operatorname{argmax}_b \{\{u_1, \ldots, u_b\} \setminus B^*\}$
end while
▷ impostazione soluzione ammissibile
for each $t = 1, \ldots, T$ **do**
    for each $b = 1, \ldots, B$ **do**
        if $b \in B^*$ **then**
            $y_{b,t} = 1$
        else
            $y_{b,t} = 0$
        end if
    end for
end for
▷ calcolo valore dell'obiettivo
$Z_p = \sum_{n=1}^{N} \sum_{t=1}^{T} w_{n,t}\, z_{n,t}$

L'obiettivo della procedura è di rendere ammissibile la soluzione prodotta dal rilassamento; per fare questo, procede spostando ambulanze da colonnine che poi non vengono effettivamente costruite ad altre che verranno realizzate. Per ricavare il valore del lower bound viene ricalcolato il valore di copertura dato dalla funzione obiettivo originaria (4.4) che sarà ovviamente più basso sia della lagrangeana (4.10) che dell'ottimo. Come si può notare dallo pseudo-codice riportato in Algoritmo 3, l'euristica procede ordinando i punti candidati per numero di utilizzi da parte di un'ambulanza. Successivamente vengono scelti i $\overline{B}$ punti che sono stati utilizzati nel maggior numero di fasce: tutti i rimanenti $B - \overline{B}$ punti candidati non possono essere costruiti e di conseguenza non è ammesso che ospitino ambulanze in nessuna fascia oraria. Riposizionate per ogni fascia le ambulanze nei punti candidati più utilizzati, viene ricalcolato il valore di copertura degli utenti per ogni t; la somma su tutte le T tipologie di fasce fornisce il valore del lower bound.

Il rilassamento lagrangeano e l'euristica sono stati utilizzati nell'*algoritmo del sottogradiente*, un algoritmo euristico di ricerca locale si sposta dalla soluzione corrente alla successiva muovendosi in direzione opposta a quella dei gradienti dei vincoli rilassati (in direzione opposta a quella di massima inammissibilità). Per la scelta dei parametri utilizzati dall'esecuzione dell'algoritmo, si sono seguiti i consigli di Beasley (in [2]). I moltiplicatori lagrangeani sono stati inizializzati a zero, il parametro di *scala* è $\pi = 1$ mentre il *passo scalare* di movimento dell'algoritmo è calcolato ad ogni iterazione con la seguente:

$$\mathcal{T} = \frac{\pi \left| \mathcal{L} - Z_p \right|}{\sum_{b=1}^{B} \sum_{t=1}^{T} G_{b,t}^2} \tag{4.11}$$

dove $\mathcal{L}$ è il valore dell'upper bound (dato dalla soluzione del rilassamento) e Z_p è quello del lower bound (dato dall'euristica lagrangeana); il denominatore è la somma di tutti i gradienti elevati al quadrato. Il valore dei gradienti dei vincoli rilassati è dato da:

$$G_{b,t} = y_{b,t} - x_b \quad \forall b \in 1, \ldots, B;\, t \in 1, \ldots, T \tag{4.12}$$

mentre ad ogni passo, i moltiplicatori lagrangeani vengono aggiornati con la seguente:

$$\mu_{b,t} = \max \left\{0,\ \mu_{b,t} + \mathcal{T} G_{b,t}\right\} \quad \forall b \in 1, \ldots, B;\, t \in 1, \ldots, T \tag{4.13}$$

La condizione di terminazione prevede tre possibilità: la prima riguarda un limite sul numero massimo di iterazioni possibili, la seconda prevede di imporre un limite minimo al valore di

scala π. Quest'ultimo viene infatti dimezzato ogni volta che si verificano troppe iterazioni (in questo caso 30) senza che venga prodotto alcun miglioramento al valore dell'upper bound: se si verifica che $\pi < 0,005$ l'esecuzione viene interrotta. L'ultima possibilità di terminazione prevede che, se i valori di lower ed upper bound convergono e finiscono per essere equivalenti, quello è il valore ottimo per il problema originario.

Risultati

Il modello è stato scritto in linguaggio *GNU MathProg* ([15]) ed il solutore utilizzato è CPLEX (versione 8.11); l'algoritmo del sottogradiente e l'euristica lagrangeana sono stati implementati in linguaggio AMPL ([10]).

I test sono stati effettuati allo scopo di verificare fino a quali dimensioni sarebbe stato possibile reggere l'esecuzione dell'algoritmo del sottogradiente; a questo scopo sono stati utilizzati dati reali riferiti alla città di Milano, prodotti dalle metodologie viste nel Capitolo 2. Il grafo utilizzato è quello completo, la rete stradale è stata modellata con $N = 12442$ nodi dove ognuno di questi è considerato un utente potenziale. La flotta comprende $V = 18$ ambulanze, la capacità di costruzione delle colonnine è di $\overline{B} = 25$ unità mentre le tipologie di fasce orarie considerate sono quelle prodotte dalla suddivisione settimanale, pari a $T = 21$. La distanza limite per la copertura dipende ovviamente dalla tipologia di fascia considerata la quale comporta un valore di velocità di circolazione dipendente dalle condizioni del traffico; i valori utilizzati vanno da un minimo di 22 km/h (distanza di copertura pari a $\overline{d} = 2,933$ km) ad un massimo di 33 km/h ($\overline{d} = 4,4$ km in 8 minuti).

Si è scelto di esaminare l'andamento dei tempi di calcolo al variare del parametro B, il numero di siti candidati individuati sul territorio. Si consideri il modello non rilassato (4.4)-(4.8): i parametri che modificati provocano la variazione del numero di vincoli sono N, T e B. Si è scelto di mantenere costante il primo considerando come utente potenziale qualsiasi punto, assegnando quindi al parametro N il più altro valore possibile sulla città di Milano; il parametro T è dipendente dalle tecniche di suddivisione in fasce viste nel Capitolo 2. La scelta è caduta sul parametro B, il numero di punti candidati ad accogliere una colonnina che, fra i tre, è di fatto è l'unico valore sotto il diretto controllo del decisore. In Tabella 4.2 sono riportati i risultati dell'esecuzione dei test. Nella colonna B sono riportati i valori utilizzati per il numero di punti candidati; la colonna ITER riporta il numero di iterazioni di sottogradiente sono state necessarie per soddisfare una delle condizioni di terminazione; nell'ultima

B	ITER	T_RUN (s)
30	1	$4,000$
40	4	$16,000$
50	5	$24,000$
100	5	$35,000$
150	3	$28,000$
200	7	$57,000$
250	26	$234,000$
300	22	$242,000$
350	32	$384,000$
400	102	$1428,000$
450	125	$1625,000$
500	137	$1918,000$

Tabella 4.2: andamento dei tempi di calcolo al variare del numero di punti candidati per il modello di ottimizzazione della copertura con costruzione dei punti di stazionamento.

colonna, la T_RUN, sono riportati i tempi complessivi impiegati dall'interprete AMPL per completare l'esecuzione di tutte le iterazioni di sottogradiente. Quest'ultimo valore, espresso in secondi, comprende sia i tempi di soluzione di lower bound ed upper bound, sia i tempi necessari all'interazione con il solutore sottostante. L'andamento dei tempi di esecuzione in funzione del numero di punti candidati è rappresentato in Figura 4.3. Dai risultati si può notare come, fino ad una dimensione di 200 punti candidati, la risoluzione di upper bound e lower bound porta i due valori a convergere molto velocemente all'ottimo: con un massimo di 7 iterazioni e 57 secondi di esecuzione complessiva i due valori si eguagliano e l'esecuzione termina con successo. Oltre la soglia dei 200 punti candidati, la risoluzione di lower ed upper bound non è più in grado di raggiungere il valore ottimo in modo immediato: il numero di iterazioni cresce fino ad arrivare ad un massimo di 137 per il caso con 500 punti candidati. È necessario specificare che, nonostante i tempi di calcolo richiesti siano più che ragionevoli per l'utilizzo previsto per l'algoritmo, i valori di B che superano il centinaio sono, a detta del decisore, assolutamente irrealistici. L'andamento dei tempi evidenzia comunque le potenzialità risolutive dell'algoritmo implementato.

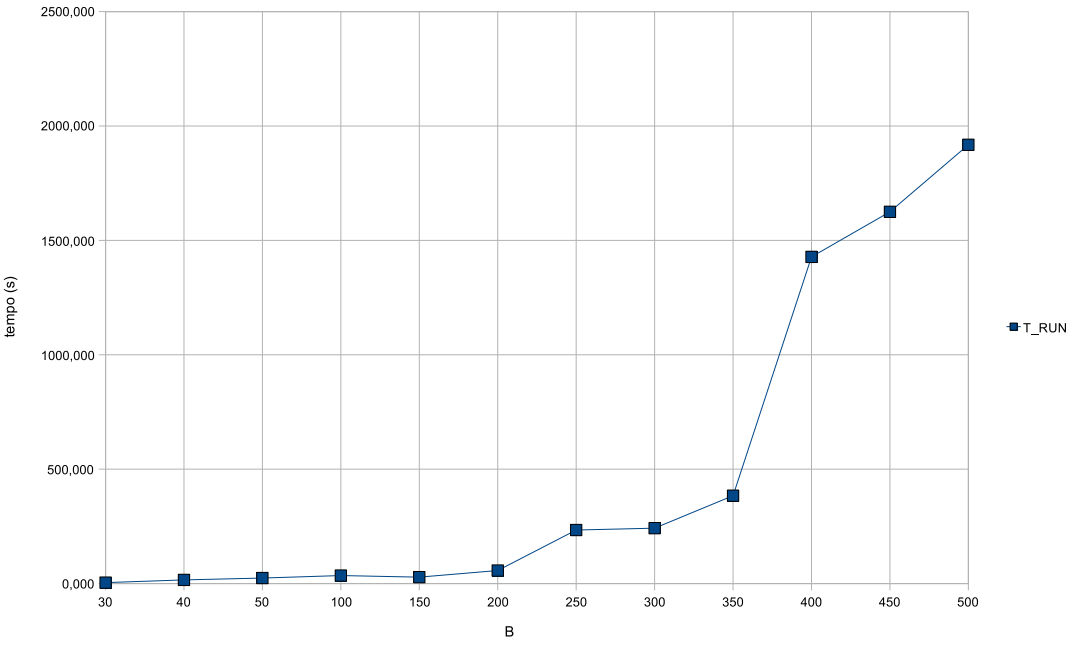


Figura 4.3: andamento dei tempi di calcolo al variare del numero di punti candidati per l'ottimizzazione della copertura con costruzione dei punti di stazionamento.

4.4 Ottimizzazione della copertura con rilocalizzazione dei mezzi

Il terzo problema preso in considerazione riguarda il posizionamento dei mezzi disponibili in modo da massimizzare la copertura e limitare il numero di rilocalizzazioni necessarie. Data la dimensione della flotta, nella fascia oraria considerata il sistema può trovarsi a dover gestire qualsiasi numero possibile di mezzi disponibili. Dato quindi il parametro N è necessario considerare altrettanti *livelli*, ognuno caratterizzato dal numero di mezzi rimasti liberi ed a disposizione; per ogni livello esiste una configurazione che, posizionando i mezzi disponibili presso i punti d'attesa, massimizza la copertura degli utenti potenziali dislocati sul territorio. Lasciando però che il modello ricavi la disposizione ottima per ognuno dei livelli in modo indipendente, si arriverebbe agli stessi risultati che produrrebbe il modello del Paragrafo 4.2. A differenza di quest'ultimo, in questo caso viene considerata la problematica legata alle

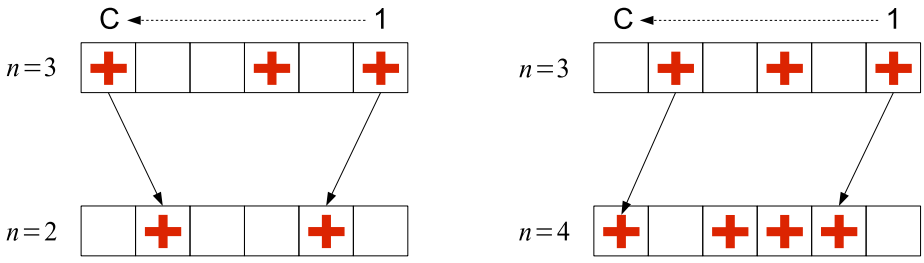


Figura 4.4: rappresentazione di alcune configurazioni ottime della disposizione dei mezzi e relative rilocalizzazioni.

rilocalizzazioni: il funzionamento del servizio prevede infatti che si possano spostare i mezzi fermi in attesa presso un'altra colonnina (si veda Figura 4.4). Potenzialmente però, passare dalla configurazione ottima di un livello a quella di un altro potrebbe richiedere di rilocalizzare anche *tutti* i mezzi disponibili. Dato che questo è chiaramente inaccettabile, è richiesto che al decisore sia data la possibilità di limitare questi spostamenti effettuati senza immediate necessità di soccorso. Grazie a questo modello viene ricavata una "catena" di configurazioni, una per ogni livello, che è possibile scorrere in entrambe le direzioni senza dover spostare più mezzi di quanti ne consideri accettabile il decisore (come mostrato in Figura 4.5). Si noti che ognuna delle configurazioni non è più singolarmente ottima ma è l'insieme di tutte le

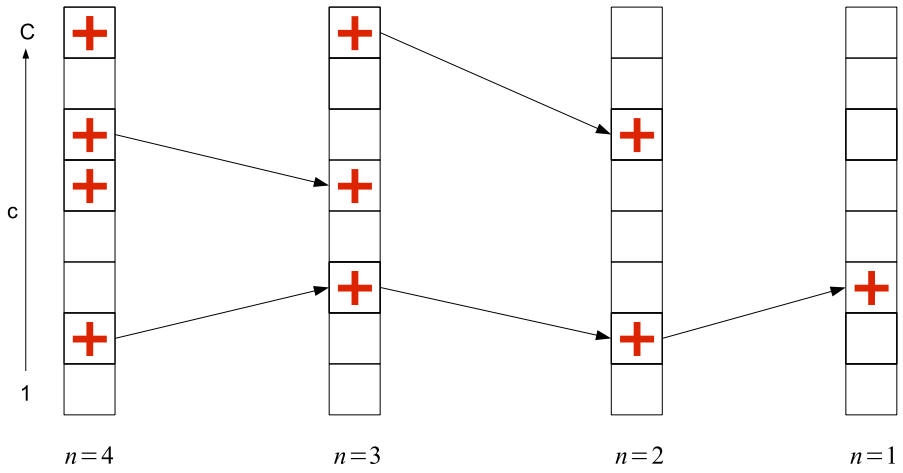


Figura 4.5: esempio di "catena" di configurazioni, una per ogni livello, con limite di rilocalizzazioni.

disposizioni per livello a fornire una copertura globalmente ottima, dato a priori il limite di rilocalizzazione.

I parametri su cui si basa il modello di ottimizzazione sono:

- $N \in \mathbb{N}^+$ numero di utenti potenziali previsti per lo scenario considerato.
- $w_n \in \mathbb{N}$ con $n = 1, \ldots, N$, vettore dei pesi associati ad ogni utente potenziale. Come spiegato per il modello di Paragrafo 4.2, il valore equivale al numero di chiamate che sono realmente state generate dal punto n.
- $C \in \mathbb{N}^+$ numero di colonnine presenti sul territorio.
- $V \in \mathbb{N}^+$ indica il numero complessivo di mezzi disponibili indipendentemente dalla natura del contratto di utilizzo.
- $\overline{d} \in \mathbb{N}^+$ distanza massima di copertura, espressa in chilometri.
- $d_{c,n} \in \mathbb{N}^2$ con $c = 1, \ldots, C$ e $n = 1, \ldots, N$, matrice delle distanze che intercorrono fra ogni colonnina c presente sull'area di competenza e qualsiasi utente potenziale n.
- $\overline{M} \in \mathbb{N}^+$ limite di rilocalizzazione. Passando da un livello ad un altro il modello può produrre una nuova configurazione dei mezzi presso i punti d'attesa atta alla massi-

mizzazione della copertura. Questo valore rappresenta quante ambulanze è possibile rilocalizzare nel passaggio tra un livello ed un altro adiacente.

Le variabili considerate da questa variante del problema di copertura sono:

- $y_{c,v} \in \{0,1\}$ con $c = 1,\ldots,C$ e $v = 1,\ldots,V$. Matrice di variabili binarie che rappresentano la presenza di un'ambulanza in ognuna delle C colonnine su tutti i possibili livelli V.
- $z_{n,v} \in \{0,1\}$ con $n = 1,\ldots,N$ e $v = 1,\ldots,V$. Matrice di variabili binarie che segnalano se ognuno degli N clienti potenziali è coperto dal servizio nella situazione in cui sono disponibili v ambulanze.

Per l'ottimizzazione della copertura con limitazione della rilocalizzazione è stato definito il seguente modello:

$$\max \quad \sum_{n=1}^{N}\sum_{v=1}^{V} w_n\, z_{n,v} \tag{4.14}$$

$$s.t. \quad \sum_{c=1}^{C} y_{c,v} = v \qquad \forall\, v = 1,\ldots,V \tag{4.15}$$

$$z_{n,v} \leq \sum_{c=1,\ldots,C\,:\,d_{n,c}\leq\overline{d}} y_{c,v} \qquad \forall\, n = 1,\ldots,N;\ v = 1,\ldots,V \tag{4.16}$$

$$\sum_{c=1}^{C} |y_{c,v} - y_{c,v-1}| \leq \overline{M} \qquad \forall\, v = \overline{M}+1,\ldots,V \tag{4.17}$$

$$y_{c,v} \in \{0,1\} \qquad \forall\, c = 1,\ldots,C;\ v = 1,\ldots,V$$

$$z_{n,v} \in \{0,1\} \qquad \forall\, n = 1,\ldots,N;\ v = 1,\ldots,V$$

Come si nota dal modello, la funzione obiettivo (4.14) massimizza il valore di copertura esattamente come avviene per le altre varianti: sommando su tutti gli utenti potenziali il prodotto fra la variabile di copertura $z_{n,v}$ ed il peso associato w_n si ottiene il valore dell'obiettivo. In questa variante però si considerano anche i vari livelli di mezzi liberi nei quali il sistema si può trovare all'interno della fascia oraria considerata. Per tenere conto di questo aspetto, il valore dell'obiettivo si ottiene effettuando la somma anche su tutti i possibili V livelli: il peso dell'utente n contribuirà solamente se egli sarà coperto quando saranno disponibili v ambulanze, quando cioè $z_{n,v}$ avrà valore uno.

Il primo vincolo, il (4.15), assicura che, per ognuno dei possibili livelli, il numero di mezzi allocati presso i punti d'attesa ne rispetti la disponibilità. Il secondo vincolo, il (4.16), controlla il valore della variabile di copertura in modo analogo a quanto visto nei modelli precedenti. Anche in questo caso, la variabile $z_{n,v}$ può avere valore maggiore di zero solo nel caso in cui si ha almeno una variabile $y_{c,v}$ a uno. La sommatoria è limitata alle colonnine entro distanza limite di servizio $\overline{d}$.

L'ultimo vincolo, il (4.17), limita il numero di mezzi che è possibile spostare di colonnina passando da un livello ad un altro. Per ognuno dei V livelli, il numero di variabili $y_{c,v}$ che differiscono dal precedente deve essere minore o uguale al limite $\overline{M}$. Per valutare quanti mezzi sono stati spostati viene calcolata la somma su tutte le colonnine della differenza (in valore assoluto) fra il valore della variabile di allocazione del livello considerato $y_{c,v}$ e la stessa riferita al livello precedente $y_{c,v-1}$. In questo modo, ognuna di queste porterà ad un valore dell'obiettivo sicuramente minore, al più uguale, rispetto a quello ricavato con il modello visto nel Paragrafo 4.2 che considera il singolo livello in modo isolato. Si noti che i valori di v partono da un minimo di $\overline{M}+1$: dal livello $\overline{M}-1$ all'$\overline{M}$ è consentito rilocalizzare anche tutti i mezzi; dal livello successivo, $\overline{M}+1$ appunto, il vincolo comincia ad essere significativo. Così come formalizzato nella (4.17), il vincolo introduce una non linearità. Per riportare il modello nell'ambito della programmazione lineare è stato necessario procedere con la sua linearizzazione:

$$\delta_{v,c} \geq y_{c,v} - y_{c,v-1} \qquad \forall\, v = \overline{M}+1, \ldots, V;\ c = 1, \ldots, C \tag{4.18}$$

$$\delta_{v,c} \geq y_{c,v-1} - y_{c,v} \qquad \forall\, v = \overline{M}+1, \ldots, V;\ c = 1, \ldots, C \tag{4.19}$$

$$\sum_{c=1}^{C} \delta_{v,c} \leq \overline{M} \qquad \forall\, v = \overline{M}+1, \ldots, V \tag{4.20}$$

$$\delta_{v,c} \in \{0,1\} \qquad \forall\, v = \overline{M}+1, \ldots, V;\ c = 1, \ldots, C$$

Il vincolo che impone l'integralità delle $\delta_{v,c}$ è di fatto ridondante in quanto questa proprietà viene dai primi due vincoli risultato della linearizzazione. In seguito però, nell'ambito della separazione in sotto-problemi, risulterà necessario imporlo; si è preferito quindi specificarlo da subito.

La tecnica di soluzione adottata

Anche in questo caso, il problema di base rimane un set covering al quale però è stata aggiunta la complicazione di dover decidere il posizionamento per ogni livello e limitare le rilocalizzazioni fra ogni configurazione e la successiva.

Anche questo problema gode della caratteristica riscontrata nel Paragrafo 4.3: eliminando i vincoli (4.18) e (4.19) è possibile ottenere un rilassamento dal quale si ricavano due problemi separati, entrambi di semplice soluzione. I vincoli eliminati vengono rilassati ed inseriti come termini di penalità nella funzione obiettivo (la (4.14)), ottenendo così la lagrangeana seguente:

$$\begin{aligned}\mathcal{L} = &\sum_{n=1}^{N}\sum_{v=1}^{V} w_n\, z_{n,v} - \\ &\sum_{v=\overline{M}+1}^{V}\sum_{c=1}^{C}\mu'_{v,c}\left(y_{c,v} - y_{c,v-1} - \delta_{v,c}\right) - \sum_{v=\overline{M}+1}^{V}\sum_{c=1}^{C}\mu''_{v,c}\left(y_{c,v-1} - y_{c,v} - \delta_{v,c}\right)\end{aligned} \tag{4.21}$$

che richiede l'introduzione di due insiemi di moltiplicatori lagrangeani, numeri reali positivi, indicati con $\mu'_{v,c}$ e $\mu''_{v,c}$:

$$\begin{aligned}\mu'_{v,c} &\geq 0 \qquad \forall\, v \in \overline{M}+1,\ldots,V;\ c \in 1,\ldots,C \\ \mu''_{v,c} &\geq 0 \qquad \forall\, v \in \overline{M}+1,\ldots,V;\ c \in 1,\ldots,C\end{aligned}$$

Il rilassamento della formulazione iniziale risulta:

$$\begin{aligned}\max\ & \sum_{n=1}^{N}\sum_{v=1}^{V} w_n\, z_{n,v} - \sum_{v=\overline{M}+1}^{V}\sum_{c=1}^{C}\left(\mu'_{v,c}\left(y_{c,v} - y_{c,v-1} - \delta_{v,c}\right) - \mu''_{v,c}\left(y_{c,v-1} - y_{c,v} - \delta_{v,c}\right)\right) \\ s.t.\ & \sum_{c=1}^{C} y_{c,v} = v && \forall\, v = 1,\ldots,V \\ & z_{n,v} \leq \sum_{c=1,\ldots,C\,:\,d_{n,c}\leq\overline{d}} y_{c,v} && \forall\, n = 1,\ldots,N;\ v = 1,\ldots,V \\ & \sum_{c=1}^{C}\delta_{v,c} \leq \overline{M} && \forall\, v = \overline{M}+1,\ldots,V \\ & y_{c,v} \in \{0,1\} && \forall\, c = 1,\ldots,C;\ v = 1,\ldots,V\end{aligned}$$

$$z_{n,v} \in \{0,1\} \qquad \forall\, n = 1,\dots,N;\ v = 1,\dots,V$$
$$\delta_{v,c} \in \{0,1\} \qquad \forall\, v = \overline{M}+1,\dots,V;\ c = 1,\dots,C$$

Come si può notare dal modello, è possibile anche in questo caso scomporlo per ottenere due sotto-problemi definiti su due insiemi disgiunti di variabili. Il primo di questi due riguarda unicamente le $y_{c,v}$ e $z_{n,v}$ e la sua formulazione è ottenuta con tutti i vincoli ed i termini della (4.21) che le contengono. Il modello risultante può essere ulteriormente scomposto, facendo in modo che possa essere risolto separatamente per ogni livello v:

$$\begin{aligned}
\max \quad & \mathcal{L}_1(v) = \sum_{n=1}^{N} w_n\, z_{n,v} - \sum_{c=1}^{C} \left(\mu'_{v,c}\,(y_{c,v} - y_{c,v-1}) - \mu''_{v,c}\,(y_{c,v-1} - y_{c,v})\right) \\
s.t. \quad & \sum_{c=1}^{C} y_{c,v} = v \\
& z_{n,v} \le \sum_{c=1,\dots,C\,:\,d_{n,c}\le\overline{d}} y_{c,v} \qquad \forall\, n = 1,\dots,N \\
& y_{c,v} \in \{0,1\} \qquad \forall\, c = 1,\dots,C \\
& z_{n,v} \in \{0,1\} \qquad \forall\, n = 1,\dots,N
\end{aligned}$$

In questo modo si ottiene il primo dei due sotto-problemi, indicato con $\mathcal{L}_1(v)$, del quale vanno risolte V istanze, ognuna in modo indipendente.

Il secondo dei due riguarda unicamente le variabili $\delta_{v,c}$ ed è ottenuto dai termini della lagrangeana (4.21) che le contengono e dal relativo vincolo di integralità:

$$\begin{aligned}
\max \quad & \mathcal{L}_2 = \sum_{v=\overline{M}+1}^{V} \sum_{c=1}^{C} \left(\mu'_{v,c}\delta_{v,c} - \mu''_{v,c}\delta_{v,c}\right) \\
s.t. \quad & \sum_{c=1}^{C} \delta_{v,c} \le \overline{M} \qquad \forall\, v = \overline{M}+1,\dots,V \\
& \delta_{v,c} \in \{0,1\} \qquad \forall\, v = \overline{M}+1,\dots,V;\ c = 1,\dots,C
\end{aligned}$$

Per ottenere il valore della lagrangeana (4.21) è necessario quindi risolvere $\mathcal{L}_2$ e V volte

$\mathcal{L}_1(v)$, per poi calcolarne la somma:

$$\mathcal{L} = \sum_{v=1}^{V} \mathcal{L}_1(v) + \mathcal{L}_2 \tag{4.22}$$

La soluzione ricavata dalla soluzione del rilassamento non porta necessariamente ad una soluzione ammissibile. Avendo rilassato il vincolo relativo alla limitazione delle rilocalizzazioni da un livello all'altro, scorrere la "catena" delle configurazioni ricavate può richiedere numerosi spostamenti. Per ottenere in modo veloce una soluzione ammissibile che rispetti anche il vincolo di rilocalizzazione, è stata definita un'euristica che si avvale di un problema ausiliario da risolvere per valori crescenti di v.

Il modello utilizzato dall'euristica è ottenuto aggiungendo alla formulazione del sotto-problema per le variabili y e z il vincolo di limitazione delle rilocalizzazioni:

$$\sum_{c=1}^{C} |y_{c,v} - y_{c,v-1}| = 1 \tag{4.23}$$

che impone la possibilità di posizionare una sola ambulanza (la v-esima) mantenendo fisse le posizioni delle $v-1$ posizionate dalle esecuzioni precedenti. Si noti che, a differenza del vincolo (4.17), questo è imposto solo sul livello v attuale.

Si consideri il modello ausiliario indicato con $\mathcal{L}'(v)$; l'algoritmo parte da $v = 1$ e, risolvendo il problema ausiliario, posiziona l'unica ambulanza disponibile in modo da massimizzare la copertura. Al passo successivo viene risolto il problema $\mathcal{L}'(v+1)$ considerando come fissa l'ambulanza posizionata al passo precedente e disponendo in modo localmente ottimo l'ambulanza aggiuntiva. L'esecuzione continua risolvendo per ogni ciclo il problema incrementando di uno il valore di v fino ad arrivare a V. In questo modo si garantisce che ogni livello differisce al massimo di una posizione dal precedente: la soluzione prodotta dall'euristica è sicuramente ammissibile.

Anche per questo problema, così come visto nel Paragrafo 4.3, i due sotto-problemi del rilassamento e l'euristica sono stati utilizzati nell'ambito dell'algoritmo del sottogradiente. I primi forniscono un upper bound all'ottimo della formulazione iniziale in quanto la soluzione ricavata non è necessariamente ammissibile; grazie all'euristica si trova velocemente una soluzione ammissibile il cui valore obiettivo funge da lower bound. Il sottogradiente è stato implementato utilizzando i valori visti nel paragrafo precedente e consigliati da Beasley ([2]).

Risultati

Il modello è stato scritto in linguaggio *GNU MathProg* ([15]) ed il solutore utilizzato è CPLEX (versione 8.11); l'algoritmo del sottogradiente e l'euristica per il calcolo del lower bound sono stati implementati in linguaggio AMPL ([10]).

I test sono stati effettuati allo scopo di valutare l'andamento dei tempi di esecuzione dell'algoritmo del sottogradiente al crescere della dimensione dell'istanza. Il grafo utilizzato è quello completo, la rete stradale è stata modellata con $N = 12442$ nodi ed ognuno di questi è un utente potenziale. La flotta comprende $V = 18$ ambulanze, la distanza limite per la copertura è stata considerata pari a $\overline{d} = 3,333$ km, corrispondente a circa 25 km/h di velocità media di percorrenza. Il limite di rilocalizzazioni è stato posto pari a $\overline{M} = 2$. Questo valore è desunto da considerazioni del decisore secondo cui, gli spostamenti dei mezzi non dovuti ad una missione di soccorso, dovrebbero essere non superiori a tale quantità. Eseguendo l'algoritmo del sottogradiente con un numero di colonnine variabile, si ottengono i tempi riportati in Tabella 4.3, il cui andamento è rappresentato in Figura 4.6. Si è scelto di utilizzare valori della variabile decisionale C che vanno da un minimo di 30 ad un massimo di 500 (colonna C). Anche in questo caso, nonostante il valore massimo sia assolutamente irrealistico, si è deciso di effettuare i test anche su istanze così grandi per mettere alla prova la capacità risolutiva dell'algoritmo implementato.

C	ITER	T_RUN (s)
30	1	3,000
40	1	3,000
50	1	3,000
100	1	7,000
150	1	12,000
200	1	25,000
250	1	38,000
300	1	21,000
350	1	24,000
400	1	66,000
450	3	271,000
500	20	2233,000

Tabella 4.3: andamento dei tempi di calcolo al variare del numero di punti d'attesa disponibili per l'ottimizzazione della copertura con rilocalizzazione dei mezzi.

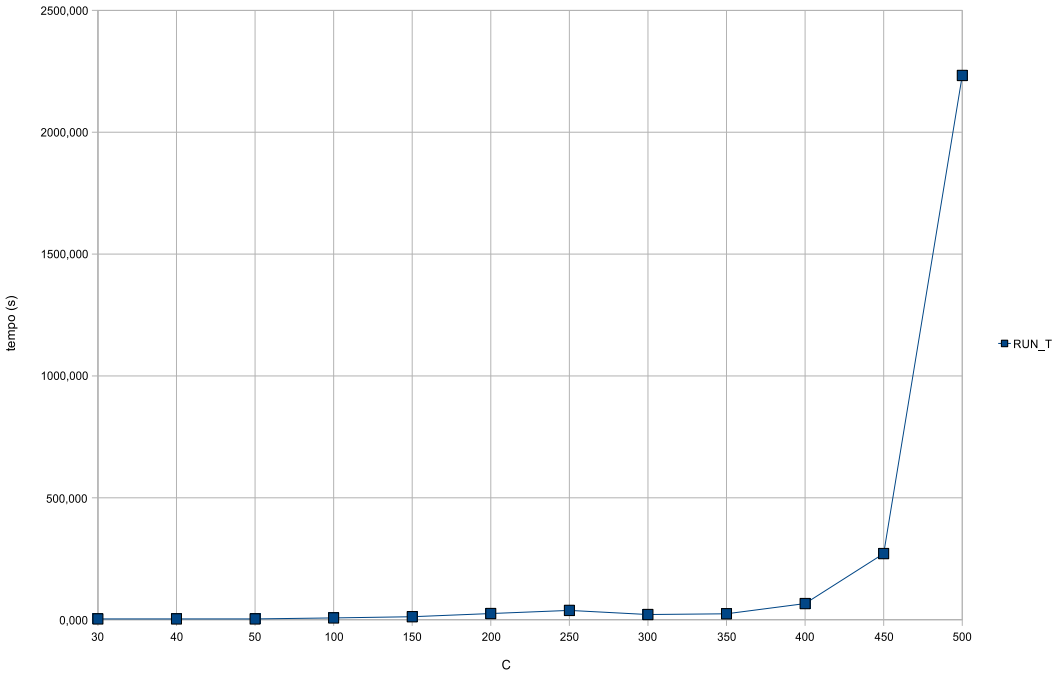


Figura 4.6: andamento dei tempi di calcolo al variare del numero di punti d’attesa disponibili per l’ottimizzazione della copertura con rilocalizzazione dei mezzi.

I dati riportati nella colonna T_RUN riguardano i tempi globali spesi dall'interprete AMPL per eseguire tutte le iterazioni di sottogradiente; questo comprende sia il tempo impiegato dal solutore, sia il sovraccarico dovuto al meccanismo di passaggio dati e risultati fra interprete e CPLEX. Nella colonna ITER è riportato il numero di iterazioni di sottogradiente completate. Come si nota dai risultati, fino ad istanze con 400 colonnine i valori di upper e lower bound convergono immediatamente e, con una singola iterazione, viene raggiunta la soluzione ottima; di conseguenza i tempi richiesti per l'esecuzione sono minimi. Nel caso delle istanze 450 e 500 colonnine non è più sufficiente una singola iterazione ma comunque, dopo un numero di cicli di sottogradiente che va da un minimo di 3 ad un massimo di 20, i valori dei bound convergono all'ottimo.

In Figura 4.7 è rappresentato l'andamento del valore di copertura al variare del limite di rilocalizzazioni $\overline{M}$; i test sono stati effettuati con $V = 20$ mezzi a disposizione e $C = 200$ colonnine disposte sul territorio. Come si può notare, il valore della funzione obiettivo cresce all'aumentare del limite; questo comportamento è sensato: meno il vincolo è stringente, più sono ampie le possibilità di rilocalizzare mezzi da un livello ad un altro. In concomitanza di $\overline{M} = 20$ è riportato il valore dell'ottimo più alto possibile dato che $\overline{M} = V$, non vengono posti vincoli e la soluzione di questa istanza è di fatto equivalente a risolvere V istanze indipendenti del problema di copertura visto nel Paragrafo 4.2. Osservando l'andamento dei valori si può notare come l'obiettivo cresca a "scalini": l'ottimo aumenta ogni qualvolta il numero di mezzi che è possibile rilocalizzare diventa sufficiente per "sbloccare" una migliore catena di configurazioni (il passaggio è evidenziato in figura dai marcatori di colore rosso).

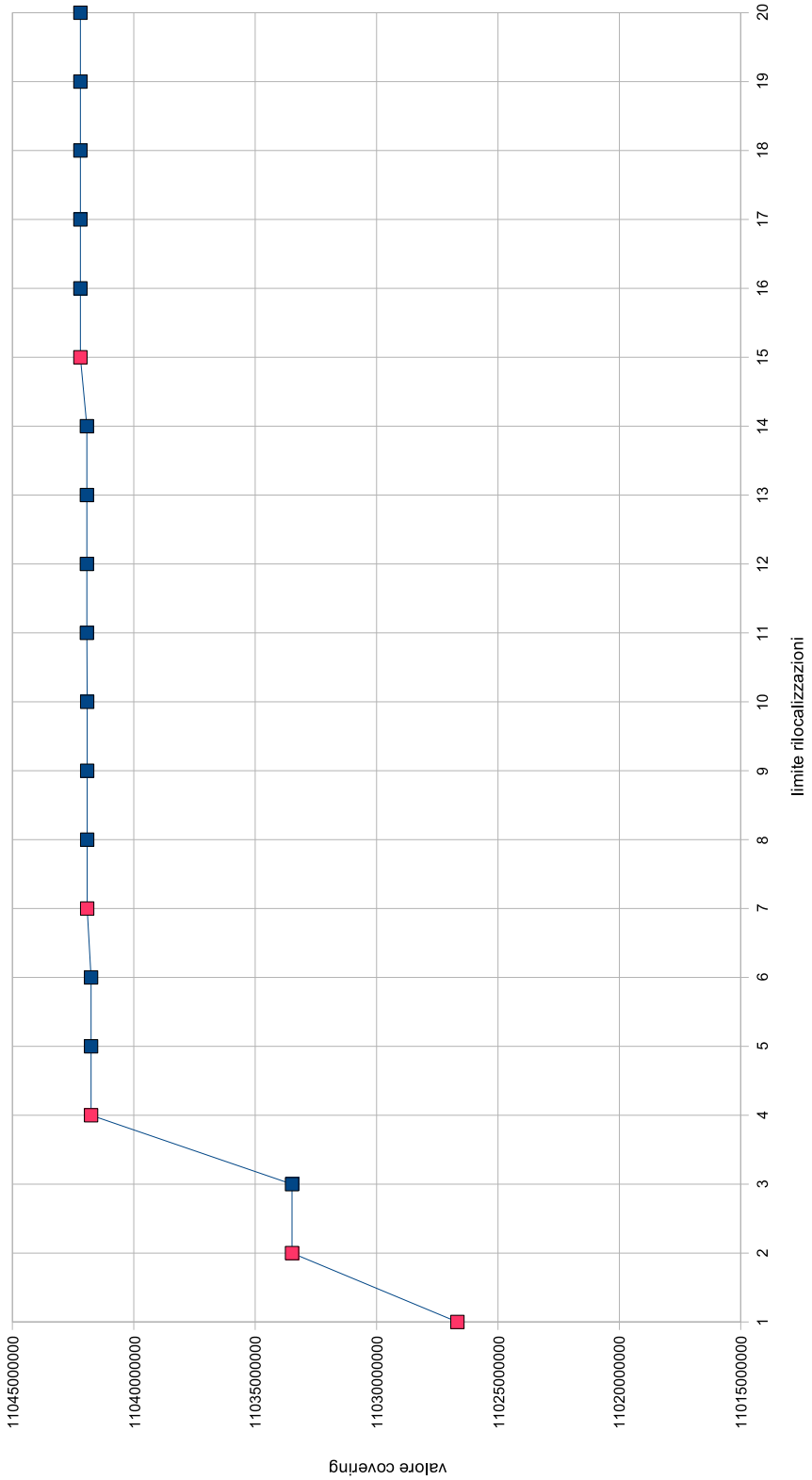


Figura 4.7: andamento del valore di copertura al variare del limite di rilocalizzazione delle ambulanze.

Capitolo 5

Conclusioni e prospettive di sviluppo

L'obiettivo prefissato per questo lavoro di tesi è quello di arrivare alla realizzazione di una serie di strumenti utili come supporto alle decisioni nell'ambito del Servizio "118". La necessità di un supporto all'analisi e pianificazione è stata espressa direttamente dai responsabili della centrale operativa della provincia di Milano; l'introduzione di strumenti numerici all'interno dei processi decisionali può essere di grande aiuto nell'affrontare problemi sia a livello tattico che strategico.

Il lavoro svolto è suddiviso in tre sezioni principali. La prima ha riguardato l'elaborazione dei dati a disposizione, al fine di caratterizzare lo scenario operativo. Nel Capitolo 2 vengono presentate le procedure realizzate a tale scopo; esse hanno richiesto lo sviluppo di strumenti appositi come una variante "stradale" e bidirezionale dell'algoritmo di Dijkstra, un algoritmo per il *geocoding* robusto ed un modello di programmazione matematica per la costruzione delle fasce temporali. La natura del Servizio "118" ha richiesto che tutto il lavoro svolto fosse a stretto contatto con il territorio di competenza. Per questo motivo è stato impiegato un Sistema Informativo Geografico (GIS) che ha permesso l'unione delle due dimensioni dei dati, quella temporale dello storico di attività e quella spaziale delle mappe geografiche.

Una volta calcolati i dati relativi agli scenari ed alle fasce temporali, è stato possibile realizzare un primo insieme di modelli. Presentati nel Capitolo 3 e basati sulla teoria dei sistemi a coda, mettono il decisore in grado di valutare l'impatto che un determinato impiego di risorse ha sulle prestazioni complessive del servizio. Questi strumenti consentono infatti di verificare i livelli di occupazione dei mezzi di soccorso, l'attesa delle chiamate non urgenti, il ricorso ad ambulanze a noleggio, il tutto a fronte di una data configurazione delle variabili decisionali; alcune di queste sono l'impiego di mezzi nella flotta, i livelli di guardia per le situazioni critiche e la tolleranza d'attesa da parte dei pazienti.

L'ultima parte del lavoro ha riguardato la definizione di modelli di programmazione lineare per risolvere problemi di posizionamento dei mezzi e realizzazione di strutture a scopo operativo. I problemi trattati nel Capitolo 4 sono definiti e risolti a livello strategico ma hanno evidenti ricadute anche in ambito tattico. Il primo aspetto considerato è stato, date le condizioni di servizio del sistema, dove posizionare i mezzi disponibili per coprire entro il tempo limite il maggior numero di clienti potenziali, ognuno pesato grazie alle analisi svolte nel Capitolo 2. La variante successiva ha previsto la definizione del problema di covering con la complicazione di massimizzare la copertura del territorio su più fasce orarie contemporaneamente; inoltre, è stato richiesto al modello di decidere, scegliendo in un insieme di luoghi candidati, dove costruire i punti di stazionamento. La complessità del problema ha richiesto

un approccio euristico, realizzato tramite l'algoritmo del sottogradiente che sfrutta due sottoproblemi ricavati dal rilassamento lagrangeano della formulazione iniziale. L'ultima variante risolve il problema del posizionamento ottimale ma, a differenza della prima versione, considera l'aspetto della rilocalizzazione dei mezzi al variare dello stato del sistema. Il modello definito ottimizza la copertura del territorio su tutti i possibili livelli di servizio, limitando le rilocalizzazioni dei mezzi necessarie per passare da una configurazione alla successiva. Anche la risoluzione di questa versione ha richiesto un approccio basato sul sottogradiente.

Gli sviluppi futuri, peraltro già preventivati e concordati con il decisore, riguarderanno l'estensione delle procedure di elaborazione dati e la loro integrazione all'interno di uno strumento completo, che permetta di trattare in modo automatizzato dati provenienti da diverse fonti in modo il più possibile semplice ed immediato. Inoltre, i modelli a coda verranno integrati e migliorati grazie al costante contatto con il decisore che provvederà, come ha fatto per tutta la durata di questo lavoro, a fornire suggerimenti, richieste e specifiche. Per quanto riguarda i modelli di copertura del territorio, gli obiettivi a maggiore priorità prevedono l'integrazione con il Sistema Informativo Geografico per la parte di visualizzazione ed interazione con l'utente. Si consideri inoltre che l'obiettivo finale per questi modelli è quello di raggiungere una completa capacità di supporto in ambito tattico, ovvero la possibilità di un impiego diretto sul campo grazie alla possibilità di ottimizzare in tempo reale la copertura del territorio controllando le rilocalizzazioni dinamiche dei mezzi.

Durante questo lavoro di tesi sono stati definiti ed implementati modelli, metodologie, procedure ed algoritmi che costituiscono una base solida e completa per un sistema di supporto ai processi decisionali. Gli obiettivi principali da perseguire nell'immediato futuro sono: l'integrazione di tutti i modelli e procedure in un unico strumento, l'estensione di questo prodotto tramite nuove funzionalità di interesse per il decisore, la progettazione e realizzazione di tutta l'interfaccia utente che permetta al decisore di utilizzare il sistema con profitto nella soluzione dei problemi legati all'attività quotidiana. Da un punto di vista globale, la direzione di sviluppo più immediata prevede la realizzazione, da portare avanti a stretto contatto con il decisore, di un sistema completo di supporto alle decisioni che sia facilmente utilizzabile dall'utente finale e permetta di sfruttare pienamente i risultati e l'incrementata capacità di visione che gli strumenti realizzati in questo lavoro offrono.

Elenco delle figure

Bibliografia

[1] Batta, R., Dolan, J. M., Krishnamurty, N. N., *The maximal expected covering location problem: Revisited*, Transportation Science 23, 277-287, 1989.

[2] Beasley, J. E., *Lagrangean Relaxation*, The Management School, Imperial College, London, May 1992.

[3] Brotcorne, L., Laporte, G., Semet, G., *Ambulance location and relocation models*, European Journal of Operational Research 147 (2003) 451-463, 2003.

[4] Burrough, P. A., *Principles of geographical information systems for land resource assessment*, Clarendon Press, Oxford, U.K, 1986.

[5] Church, R. L., ReVelle, C. S., *The maximal covering location problem*, Papers of the Regional Sciences Association 32, 101-118, 1974.

[6] Daskin, M. S., Stern, E. H., *A hierarchical objective set covering model for emergency medical service vehicle deployment*, Transportation Science 15, 137-152, 1981.

[7] Daskin, M. S., *A maximum expected location model: Formulation, properties and heuristic solution*, Transportation Science 7, 48-70, 1983.

[8] Dijkstra, E. W., *A note on two problems in connection with graphs*, Numer. Math. 1, 269-271, 1959.

[9] Eaton, D. D., Daskin, M. S., Simmons, D., Bulloch, B., Jansma, G., *Determining emergency medical deployment in Austin, Texas*, Interfaces 15 (1), 96-108, 1985.

[10] Fourer, R., Gay, D. M., Kernighan, Brian W., *AMPL: A Modeling Language for Mathematical Programming*, Duxbury Press, Brooks, Cole Publishing Company, 2002.

[11] Fujiwara, O., Makjamroen, T., Gupta, K. K., *Ambulance deployment analysis: A case study of Bangkok*, European Journal Of Operational Research 31, 9-18.

[12] Gendreau, M., Laporte, G., Semet, F., *Solving an ambulance location problem by Tabu search*, Location Science 5, 75-88, 1997.

[13] Gendreau, M., Laporte, G., Semet, F., *A dynamic model and parallel Tabu search heuristic for real-time ambulance relocation*, Parallel Computing 27, 1641-1653, 2001.

[14] Goldberg, J., Dietrich, R., Chem, J. M., Mitwasi, M. G., *Validating and applying a model for locating emergency medical services in Tucson, Arizona*, European Journal Of Operational Research 49, 308-324, 1990.

[15] *GNU Linear Programming Kit, Modeling Language GNU MathProg, Version 4.9*, Free Software Foundation, Gennaio 2006.

[16] Hillier, F. S., Lieberman, G. J., *Introduction to Operations Research*, McGraw-Hill, Industrial Engineering Series, 1995.

[17] Hogan, K., ReVelle, C. S., *Concepts and applications of backup coverage*, Management Science 34, 1434-1444, 1986.

[18] Larson, R. C., *Approximating the Performance of Urban Emergency Service Systems*, Operations Research 23 (5), 845-867, 1975.

[19] Larson, R. C., *A hypercube queueing model for facility location and redistricting in urban emergency services*, Computers and Operations Research 23, 845-868, 1974.

[20] Little, J. D. C., *A Proof for the Queueing Formula:* $L = \lambda W$, Operations Research, 9 (3), 383-387, 1961.

[21] Marianov, V., ReVelle, C. S., *The capacitated standard response fire protection siting problem: deterministic and probabilistic models*, Annals of Operations Research 40, 303-322, 1992.

[22] Repede, J. F., Bernardo, J. J., *Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky*, European Journal Of Operational Research 75, 567-581, 1994.

[23] ReVelle, C. S., Hogan, K., *The maximum availability location problem*, Transportation Science 23, 192-200, 1989.

[24] Schilling, D. A., Elzinga, D. J., Cohon, J., Church, R. L., ReVelle, C. S., *The TEAM/FLEET models for simultaneous facility and equipment sitting*, Transportation Science 13, 163-175, 1979.

[25] Staff 118, *I diciotto telefoni che salvano la vita*, disponibile presso `www.118milano.it` (consultato il 27/03/2007).

[26] Staff 118, *Protocolli Operativi Centrale Operativa 118 di Milano*, PRO.E.CO.004, disponibile presso `www.118milano.it` (consultato il 27/03/2007).

[27] Toregas, C. R., Swain, R., ReVelle, C. S., Bergman, L., *The location of emergency service facilities*, Operations Research 19, 1363-1373, 1971.

Si ringraziano:
il relatore Prof. Giovanni Righini, per il supporto costante, attento e competente;
il correlatore Dott. Giovanni Sesana, responsabile S.S.U.Em. 118 Milano presso l'Azienda Ospedaliera Ca' Granda Niguarda, per la fiducia accordatami e la gentile disponibilità;
i Dottori Andrea Bettinelli, Alberto Ceselli ed Andrea Pinciroli per l'aiuto prezioso ed indispensabile;
la mia Famiglia, per esserci.